网络舆情分析技术

蔡皖东　编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

随着互联网技术的快速发展和广泛应用，网络媒体已经成为重要的信息传播和交流平台，同时也是网络舆情形成和传播的主要载体。网络舆情通常由突发社会公共事件触发，反映了人们对某一公共事件所表达的认知、态度、情感和倾向性，具有虚拟化、快捷化、多元化、开放性、匿名性及互动性等特点。随着网络舆论影响力的增强，网络舆情已经成为各级政府了解社情民意的重要窗口。

随着网络舆论对社会和公众影响的不断增大，出现了网络炒作、造谣传谣等不良的现象，损害了网络媒体公信力，扰乱了网络正常传播秩序，产生了错误的舆论导向，极易引发社会群体性事件。因此，加强互联网管理和舆论治理非常重要和必要。

国家大力推进网络舆情监控体系建设，加强对网络舆情监测和引导。网络舆情监测系统在互联网舆情监测中发挥了重要的作用，其系统核心技术就是网络舆情分析技术。网络舆情分析技术主要涉及网络信息采集技术、网络舆情传播机制、话题检测与跟踪技术、文本分割技术、文本情感分析技术等。

本书主要介绍网络舆情分析所涉及的主要方法和关键技术，全书共分 7 章，分别介绍网络舆情概论、网络信息采集技术、微博网络信息传播机制、网络论坛舆情传播机制、话题检测与跟踪技术、文本分割技术和文本情感分析技术。在介绍主要模型和算法时，还给出了模型和算法的实验验证，以便读者加深对模型和算法的理解。

本书可作为网络空间安全学科相关专业的研究生和本科生教材，也可作为从事相关工作的科技人员及管理人员的参考书。

图书在版编目（CIP）数据

网络舆情分析技术 / 蔡皖东编著. — 北京：电子工业出版社，2018.2
ISBN 978-7-121-33354-5

Ⅰ. ①网… Ⅱ. ①蔡… Ⅲ. ①互联网络－舆论－研究 Ⅳ. ①G206.2

中国版本图书馆 CIP 数据核字（2017）第 320373 号

策划编辑：窦 昊
责任编辑：窦 昊
印 刷：
装 订：
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编：100036
开 本：787×980 1/16 印张：14 字数：323 千字
版 次：2018 年 2 月第 1 版
印 次：2018 年 2 月第 1 次印刷
定 价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010) 88254466，douhao@phei.com.cn。

前　言

随着互联网技术的快速发展和广泛应用，网络媒体已经成为重要的信息传播和交流平台，网络媒体被称为继报纸、杂志、广播、电视四大传统媒体之后的第五媒体，人们通过网络媒体提供的新闻、微博、论坛、博客等网络服务可以快捷地交流和共享信息资源，实现广泛的沟通交流，受众不仅是信息的接收者，也是信息的传播者。因此，网络媒体成为民众表达民意、交流情感的主要窗口，同时也是网络舆情形成和传播的载体。

网络舆情通常由突发社会公共事件触发，反映了人们对某一公共事件所表达的认知、态度、情感和倾向性，具有虚拟化、快捷化、多元化、开放性、匿名性及互动性等特点，主要通过网络新闻、微博、论坛帖子、博客博文等网络媒体进行传播，其中微博网站和网络论坛是主要的信息传播平台。随着网络舆论影响力的增强，网络舆情已经成为各级政府了解社情民意、改进工作作风、提高执政能力的重要窗口。

随着网络舆论对社会和公众影响的不断增大，出现了以网络炒作为营生的网络公关公司、网络推手、网络水军等。网络公关公司以营利为目的，为了炒作某个话题、人物或产品，雇用了大量的网络水军，在网络推手的组织下，短时期内在互联网的各大网络论坛上大量地发布煽动性帖子，并通过跟帖、转帖及编发评论等方式炮制网络热点事件，营造虚假民意，从而形成虚假的网络舆情。网络公关公司、网络推手、网络水军等形成了灰色利益链，他们在实现“客户”利益的同时也获得自身利益。随着微博的兴起，网络水军把造谣和传谣的主要阵地从网络论坛转至微博，形成了网络推手、网络水军造势与网络大 V 的转发影响力相结合的网络谣言制造和传播的灰色利益链，个别网络大 V 在突发重大公共事件中发表不当言论，或者以“求辟谣”“求证”等方式故意扩散谣言，导致不明真相的网民盲目跟风，损害了网络媒体公信力，扰乱了网络正常传播秩序，产生了错误的舆论导向，危及政府的公信力，极易引发社会群体性事件。

中央高度重视互联网管理和舆论治理，在《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》中指出，应该“牢牢把握正确舆论导向，健全社会舆情引导机制，传播正能量。加强网上思想文化阵地建设，实施网络内容建设工程，发展积极向上的网络文化，净化网络环境”。国家通过开展打击网络谣言等专项行动，依法惩戒了利用互联网进行造谣惑众的“秦火火”“立二拆四”等网络大 V，使不法网络大 V 和网络谣言遭到压制性

打击，一些被称为“推墙派”“凿船派”“体制婊”的网络负能量群体被有效遏制，网络舆论空间逐步呈现风清气正之势。

由于网络舆情已成为各级政府了解社情民意的重要窗口，因此国家大力推进网络舆情监控体系建设，各级宣传主管部门以及主流新闻媒体等都设立了网络舆情监测机构，加强对网络舆情监测和引导。在网络舆情监测中，面对海量的网络信息，必须借助信息技术实现网络舆情监测的自动化和智能化。于是，各种网络舆情监测系统便应运而生，系统的核心技术就是网络舆情分析技术。

网络舆情分析技术是一种大数据应用，首先需要对网络舆情来源的网络信息进行实时监测和采集，然后对采集到的网络信息进行处理和分析，主要涉及网络信息采集技术、网络舆情传播机制、话题检测与跟踪技术、文本分割技术、文本情感分析技术等。网络舆情分析技术属于交叉性技术，涉及自然语言处理、计算语言学、人工智能、机器学习、信息检索、数据挖掘等多个研究领域和学科方向。

本书主要介绍网络舆情分析所涉及的主要方法和关键技术，全书共有 7 章，第 1 章为网络舆情概论，主要介绍网络舆情概述、网络舆论空间治理、网络舆情传播平台、网络舆情分析技术等内容；第 2 章为网络信息采集技术，主要介绍搜索引擎概念、网络蜘蛛概念、网页搜索算法、网页相似度计算、主题蜘蛛组成等内容；第 3 章为微博网络信息传播机制，主要介绍微博用户转发特性、微博转发行为预测、微博转发峰值分析、微博意见领袖识别等内容；第 4 章为网络论坛舆情传播机制，主要介绍网络论坛舆情形成模型、网络论坛意见领袖识别、网络水军热帖检测、网络水军账号检测等内容；第 5 章为话题检测与跟踪技术，主要介绍话题检测与跟踪的基本概念、相关技术、话题检测算法、话题跟踪算法、热点话题检测等内容；第 6 章为文本分割技术，主要介绍文本分割的基本概念、基于 LDA 模型的文本分割、基于 VSM 模型的文本分割等内容；第 7 章为文本情感分析技术，主要介绍文本情感分析的基本概念、句子情感分析方法、段落情感分析方法、文本情感分析模型等内容。

本书在介绍主要模型和算法时，给出了模型和算法的实验验证，以便读者加深对相关模型和算法的理解。本书可以作为网络空间安全学科相关专业的研究生和本科生教材，对从事相关工作的科技人员及管理人员也能起到很好的参考和启示作用。

由于网络舆情分析技术比较复杂，本书很难覆盖相关技术的方方面面，难免存在不足和疏漏之处，欢迎广大读者批评指正。

本书的主要内容来自于作者及其研究团队的工作总结，张博、罗知林、徐会杰、石磊、杨惠、樊娜及彭冬等同志参与了相关研究工作，并为本书的撰写做出了贡献，对此表示感谢。

最后，感谢西北工业大学教材专著出版基金对本书的大力资助。

作　者

于西北工业大学

目　录

第 1 章

网络舆情概论

1.1 网络舆情概述

随着互联网技术的快速发展和广泛应用，网络媒体已经成为重要的信息传播和交流平台，网络媒体被称为继报纸、杂志、广播、电视等四大传统媒体之后的第五媒体，人们通过网络媒体提供的新闻、微博、论坛、博客等网络服务可以快捷地交流和共享信息资源，实现广泛的沟通交流，受众不仅仅是信息的接收者，同时也是信息的传播者。因此，网络媒体成为民众表达民意、交流情感的主要窗口，同时也是网络舆情传播的载体。网络舆情通常由突发社会公共事件触发，反映了人们对某一公共事件所表达的认知、态度、情感和倾向性，主要通过网络新闻、微博、论坛帖子、博客博文等网络媒体进行传播，其中微博网站和网络论坛是主要的信息传播平台。随着网络舆论影响力的增强，网络舆情已经成为各级政府了解社情民意、改进工作作风、提高执政能力的重要窗口。

下面简单介绍网络舆情基本概念、网络舆情演化过程以及网络舆情实例等。

1.1.1 舆情与舆论

“舆”的含义是民众或公众，“情”的含义是情绪或意愿，“舆情”的含义是公众的情感或情绪。“舆情”一词最早出现在《旧唐书》中。唐昭宗在乾宁四年（公元 897 年）的一封诏书中称：“朕采于群议，询彼舆情，有冀小康，遂等大用”。“舆”与“情”两字的连用，最初是指百姓的情感、情绪。现在《新华字典》中也采用了这个解释。在《辞源》中，则把“舆情”解释为“民众的意愿”。

由于舆情问题涉及社会学、心理学、新闻传播学、政治学等相关领域，目前对舆情还没有一个权威、统一的定义，不同领域的学者分别从不同角度对舆情概念进行了诠释。

从社会学的角度，将舆情定义为“在一定的社会空间内，围绕中介性社会事项的发生、发展和变化，作为主体的民众对作为客体的国家管理者产生和持有的社会政治态度”[1]。

从社会心理学的角度，将舆情定义为“由个人及各种社会群体构成的公众，在一定历

史阶段和社会空间内，对自己关心或与自身利益紧密相关的各种公众事务所持有的各种情绪、意愿、态度和意见交错的总和”[2]。

从舆论引导工作的角度，将舆情解释为“在一定的社会空间内，围绕特定的舆情因变事项的发生、发展和变化，在民众中产生和存在的对执政者及其所持有的政治价值取向的社会政治态度。或简述为，舆情是舆情因变事项发生、发展和变化过程中，民众所持有的社会政治态度”[3]。

舆情与舆论两个词一字之差，很容易混淆。实际上，两者的概念既有联系又有区别。

广义的舆论是指人们的认知、态度、情感和行为倾向的集聚表现，是多数人形成的一致共同意见，也就是持有某种认知、态度、情感和行为倾向的人群需要达到一定的数量，否则不能称为舆论。而舆情是指人们的认知、态度、情感和行为倾向的原始表露，可以是零散的、非体系化的，也不需要得到多数人的认同，是多种不同意见的简单集合，这也是最容易将两者混淆的地方。舆情聚集时有可能转化为舆论，通过舆论引导可以使舆情转化为良性舆论。

舆情和舆论都表现为公众的意见、情绪和态度，舆情虽然是多种意见、情绪和态度交织的总和，但是其中包含着小范围多数人的意见，就是舆论。当这种意见被社会大多数人认同的时候，就会转换为声势浩大的社会舆论。当正确和公正的意见被广大群众推崇和追随时，代表着历史发展的必然趋势，从而形成了主流民意。针对某一公共事务的分散和错综复杂的舆情，向一致有序的舆论与民意的转化是一种必然趋势。

1.1.2 网络舆情

网络舆情是指通过互联网表达和传播的舆情，反映了人们对某一公共事件所表达的认知、态度、情感和倾向性，具有虚拟化、快捷化、多元化、开放性、匿名性及互动性等特点。

网络舆情通常由突发社会公共事件触发，突发社会公共事件包括自然灾害、重大事故、公共卫生、社会安全、媒体事件、地方经济、社会治理、官吏腐败等多个方面，推动网络舆情形成和传播的网络媒体主要是网络新闻、微博、论坛帖子、博客博文等，其中微博网站和网络论坛是网络舆情形成和传播的主要信息传播平台。

网络舆情作为舆情的网络表现形式，具有如下特点：

（1）自发性。在互联网上，人人都可以自由、自发地发表意见和表达态度，每个人既是信息的发布者，又是信息的评论者，同时还是信息的传播者。任何一个公共事件发生后，网民都会自发地通过微博、论坛帖子、博客博文等网络媒体自由发表意见，表达自己的观点、情绪和态度。因此，网络舆情能够比较客观地反映现实社会的矛盾和公众的诉求。

（2）多元性。在虚拟的网络空间中，人们不再像现实生活中总是掩饰自己真实的想法和感受，而是更愿意表达自己真实的意见、情绪和态度，能够真实地反映人们不同的思想形

态、文化观念、价值取向、生活准则以及道德规范等。因此，网络舆情在价值传递、利益诉求等方面呈现出多元性特点。

（3）时效性。由于网络媒体打破了时间和空间上的界限，网民可以随时随地在网上发表意见，网络舆情的形成非常迅速。当一个公共事件发生后，网民可以立即在网上发表意见，网民意见由点到面，由散到聚，迅速汇聚成公共意见，形成强大的意见声势。另一方面，随着其他社会热点事件的发生，一个舆情事件很快被新涌现的公共事件所掩盖，持续时间通常为一周左右，呈现较强的时效性。

（4）偏差性。社情民意是基于网络舆情中最普遍、最尖锐的一部分，包含了广大网民不同的利益诉求，在一定程度上真实反映了他们的意见和呼声，但网络舆情不能完全等同于所有社会群体的立场。虽然网民在网络媒体上可以自由地表达自己的观点和情绪，但是网民的言论及其影响与所承担的社会责任是脱节的，网民随意发表某些不负责任的言论甚至谣言，导致网络舆情表达失真，与真实的社会舆情存在偏差，极易引起网民情绪走向极端，导致“网络暴力”频现。

（5）从众性。在网络舆情形成过程中，一些网民并不直接发布信息，而是通过关注和转发他人的信息来表达自己的态度和倾向性，特别是意见领袖发布的信息。意见领袖是指在信息传播网络中经常发表意见并具有相当影响力的“活跃分子”，在意见领袖的引导和影响下，网民通过微博转发、论坛跟帖等方式，推动网络舆情的形成和发展，局部意见可能演化成网络舆情。

1.1.3　网络舆情演化

当一个突发社会公共事件发生后，随着事件的发展以及时间的推移，网络舆情经历一个形成期、高涨期、波动期和消退期的演化过程，而事件的处置措施和应对能力对网络舆情的发展和走势起到至关重要的作用。

（1）舆情形成期。舆情形成期也称为发酵期，当一个突发社会公共事件发生后，经过传统媒体或网络媒体的报道，引起网民的关注和热议，通过微博、论坛帖子等方式表达自己的意见、情绪和态度，并在互联网上不断地传播，经过不断地发酵而成为被广泛关注的热点事件，形成了网络舆情。在发酵期，如果事件处置和应对得当，则有可能及时化解尚未形成的舆论热点。

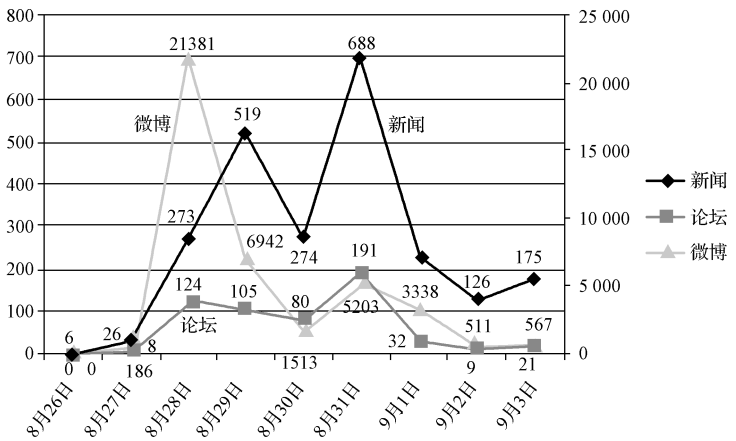
（2）舆情高涨期。网络舆情形成后，随着网民的情绪等不断高涨，热点事件的关注度越来越高，不断吸引更多网民的关注，其关注度和影响力达到高峰。根据事件性质和发展，高涨期持续时间长短不一。在高涨期，如果事件处置和应对得当，则有利于消解高涨的网络舆论，释放公众的情绪。

（3）舆情波动期。网络舆情达到高峰后，随着事件的发展和处置，网络舆情进入波动期，呈现出波浪式发展态势，在事件处置不当或出现新情况时，网络舆情可能出现又一

波高峰。在波动期，良好的处置措施和应对能力能够使网络舆论得到有效的化解，进入舆情消退期。

（4）舆情消退期。波动期持续一段时间后，由于事件的处理、新事件的出现等原因，使人们的关注点发生了转移，转向新的目标。这时，原来的关注热点逐渐变冷，网络舆情进入消退期，最终沉寂。在消退期，有效的应对措施能够起到形象修复和再造的作用。

图 1-1 为杨达才“微笑-名表门”事件的网络舆情关注度走势图、其中，纵坐标为新闻/帖子/微博的数量，单位为条数；横坐标为日期。从中可以看出微博、论坛、新闻等网络媒体的网络舆情演化过程。



杨达才“微笑-名表门”舆情关注度走势，数据来源：人民网网络舆情监测室[4]

图 1-1　网络舆情演化过程示例

从图 1-1 还可以看出，作为一个特定事件的网络舆情，应具有如下特征：

（1）关注度高。在短时间内，针对该事件所发布或转发的新闻、微博、论坛帖子等各自数量（条数）达到一定的数量级，说明该事件引起社会和网民的广泛关注，成为关注度较高的热点事件。例如，在图 1-1 中，新闻的高峰值为 688 条、微博的高峰值为 21384 条、论坛帖子的高峰值为 191 条。并且新闻、微博、论坛帖子的数量之间存在潜在的关联性。

（2）持续时间长。关注该事件的网络舆情从形成到消退所持续的时间（天数）比较长，一般为一周左右。例如，在图 1-1 中，该事件的网络舆情持续时间超过了 9 天，持续时间较长。

（3）倾向性多元化。在针对该事件的所有新闻、微博、帖子中，对该事件所持有的情感倾向性呈现多元化，包括正面或负面、赞扬或批评、支持或质疑、肯定或否定等。

因此，一个突发社会公共事件的网络舆情应当具有很高的关注度，其关注度应持续较长的时间，并且对该事件的情感倾向性通常表现为多元化。表 1-1 为近年来由突发社会公共

事件触发的网络舆情情况，突发社会公共事件包括自然灾害、重大事故、公共卫生、社会安全、媒体事件、地方经济、社会治理、官吏腐败等各个方面。数据来源于人民网网络舆情监测室[4]。

表 1-1　代表性的自然灾害类网络舆情

事件名称	发生时间	关注度高峰值（条数）		情感倾向性（%）		
		新闻	微博	负面	正面	其他
浙江余姚市“菲特”台风灾害	2013 年 10 月 8 日	8 011	69 482			
黑龙江林甸中储粮库火灾事件	2013 年 5 月 31 日	6527	209 700	72	21	7
中石油输油管道爆炸事件	2013 年 11 月 22 日	5 472	82 704			
上海外滩踩踏事件	2014 年 12 月 31 日	12 843	100 882			
河南鲁山县老年公寓特大火灾事件	2015 年 5 月 25 日	6 310	9 529	77	20	3
上海黄浦江死猪飘浮事件	2013 年 3 月 9 日	2 746	315			
甘肃兰州水污染事件	2013 年 4 月 11 日	2 780	41 200	73	24	3
浙江温岭患者杀医生事件	2013 年 10 月 29 日	6 194	115 092			
山东招远故意杀人事件	2014 年 5 月 28 日	1 979	1 979	40	57	3
黑龙江嫌疑犯杀警越狱事件	2014 年 9 月 2 日	1 928		60	40	
快播公司涉黄视频被查事件	2014 年 4 月 16 日	1 010	2 660	47	49	4
红十字会赈灾送棉被事件	2014 年 7 月 19 日	1 198	74 307	69	23	8
21 世纪网涉嫌敲诈事件	2014 年 9 月 3 日	1 445	2 850	14	85	1
呼格吉图冤假错案事件	2014 年 10 月 30 日	400	100 000		94	6
广东东莞扫黄案事件	2014 年 2 月 9 日	6 800	15 956		95	5
湖南临武瓜农死亡事件	2013 年 7 月 17 日	3 738	378 900	92		8
云南导游辱骂游客事件	2015 年 5 月 1 日	3 058	4 678	50	41	9
优衣库试衣间不雅视频事件	2015 年 7 月 15 日	4 030	6 000	17	78	5
杨达才“微笑-名表门”事件	2012 年 8 月 26 日	688	21 384		93	7
湖南万宁校长带女生开房事件	2013 年 5 月 13 日	1 621	1 218 000	18	82	

1.1.4　网络舆情实例

发表在求是网的“2015 年中国互联网舆情研究报告”中，作者小兵章嘎以舆情观察者的角度，根据事件重要性、影响力、长远意义等三个方面，列选出 2015 年中国十大互联网舆情。该文属于其个人观点，摘抄如下。

1. “穹顶之下”，人为操控舆情的经典案例

2015 年全国“两会”前夕，由原央视节目主持人柴静制作并担任主持的《穹顶之下》纪录片，以迅雷不及掩耳之势强势占领各大商业网站、新媒体头版头条，甚至吸引众多官方媒体、主流媒体跟风转载，成为上至最高领导、环保部长，下至平民百姓共同关注、讨论的话题，“全民刷屏”、“全民热议”、“全民转载”，各大媒体、平台在 24 小时之内几无

招架还手之力。“雾霾”团队以“一己之力”调用了近半数网络媒体、平台、大 V 等资源，成功实现“全民走心”，让环境问题再度成为中国老百姓抹不去的伤痛，柴静本人也因本片成功而入选美国《外交政策》全球百大思想者。《穹顶之下》的议题策划能力，已经远远超出一般省市级媒体的舆论掌控能力。

2．“长江沉船”，舆情防御的经典案例

2015 年 6 月 1 日晚，在长江中游发生了“东方之星”游轮倾覆沉船事故。如此重大的突发公共事件引起了全社会的共同关注，中国各大媒体、社会舆论场、自媒体智库、网络舆论场，以及境外媒体和网民集体聚焦长江沉船事故的跟踪报道。但与云南鲁甸地震、上海外滩踩踏、马航失联、漳州古雷爆炸、台湾复兴坠机、尼泊尔地震、贵州贵阳塌楼等重大灾难性事件相比，“6·1 长江沉船事件”呈现出迥然不同的特点，官方的准确发布、正义网民的精准发力，使得舆情“出奇的平静”。从舆论施压角度来说，并没有一个实体组织、政府部门、政府官员、游船负责人成为媒体或网络共同口诛笔伐的焦点，对于如此重大的突发事件，网络能够有如此的表现，全然得益于官方和正义网民的上下齐手、协力奋战。

3．“天津爆炸”，舆情次生灾害的经典案例

2015 年 8 月 12 日，在天津市滨海新区天津港危险品仓库发生了特大火灾爆炸事故。“8·12 天津爆炸事故”舆情之复杂、舆论“战斗”之激烈，历年罕见。与“7·23 甬温线动车事故”重大舆情类似，新闻发布是处置“天津爆炸”舆情的咽喉所在，及时掌握媒体运行规律，善于应对，就能化险为夷；如若不懂媒体规则、不了解社会心理而采取错误不明智的举动，不但无助于解决问题，反而会成为舆情应对的“自残行为”。显然，“8·12 天津爆炸事故”在新闻发布这个关键点是失位的，“快报事实、慎报原因”这一重要舆情应对法则，在该舆情处置中被还给了“课堂”。

4．“庆安枪案”，死磕派律师公知的倾巢之战

2015 年 5 月 2 日，在黑龙江省绥化市庆安县火车站候车厅，执勤民警开枪击毙涉嫌暴力袭警的当地农民，这一事件引起了社会高度关注，网上出现了各种质疑和争议。在该舆情中，“@李方平律师”、“@李长青律师”等公知边缘账号率先发力，“@徐昕”、“@袁欲来律师”等核心律师大 V 在舆情关键点准确施压，成功制造出一个有利于公知的巨大舆论阴影。所幸，警方吸取了王文军案件的深刻教训，线下执法并未受到舆论施压的影响。

5．“肃宁枪击”，舆情转弯的经典案例

2015 年 6 月 8 日深夜，在河北省肃宁县发生了特大枪击案，引起了社会高度关注。“肃宁枪击”舆情紧跟着“长江沉船”事件而来，央视节目主持人白岩松的新闻评论“是什

么让一个 50 多岁的老汉端起了他的猎枪”，成为了微博舆论场发酵的笑柄，究其缘由，该言论与网络舆论场的公知言论如出一辙。舆论假象在白岩松的口中成为了现实，虚假民意的再次投射给了权威媒体，白岩松不幸成为了“舆情中枪”的对象。

6．“老毕视频”，网络舆论秒杀明星光环

2015 年 4 月初，“老毕视频”出现在互联网上，不断被转发和评论，成为被社会广泛关注的热点事件。与前些年舆论场被核心媒体和明星人物所掌控相比，“老毕视频”事件标志着舆论话语权开始与核心力量剥离，草根中 V、小 V 开始真正拥有了舆论话语权和公信力。从另一个角度来看，“老毕视频”事件充分说明了一切的舆论“权威”只是漂浮在云端的虚幻，互联网的“扁平”特性已经在中国舆论场得到充分的体现。

7．“优衣库门”，恶意营销制造低俗的狂欢

2015 年 7 月中旬，优衣库产品广告视频在微博中热传，视频中含有“涉黄”内容，被称为“优衣库门”事件。对“涉黄”和美女信息的猎奇，使得该营销舆情炒作地尤为成功。但从管理角度来说，微博、微信圈的私人信息保护属性，在某些时候容易成为舆情发酵的不确定因素，官方对半封闭圈的监管难以到位，成为互联网管理的一大安全隐患，进一步加强互联网半私密空间的监管，已经刻不容缓。

8．“38 元大虾”，线下线上舆情同步的经典案例

2015 年国庆长假，青岛“38 元大虾”舆情突然席卷全国，在重大事件和舆情的空当，“大虾”成为了媒体和网络最火的名词，随后而来的“31 米蟹绳”、“25 元清水鱼”等“舆情搭车”事件，使得中国旅游景区价格乱象成为老百姓和网络共同热议的话题，也成为某些不良（无知）媒体舆情乱炒的题材。

9．“加多宝辱英烈”，资本力量操控舆论的首次倾覆

2015 年 4 月 17 日，在加多宝公司官方微博发布的凉茶营销广告中使用了有辱英烈邱少云、赖宁的段子，激起了网友的强烈抗议。加多宝在该舆情中表现出典型的“懂网却不会用网”、“懂舆论却不会顺应舆论”等资本通病，导致该事件的后续发展完全背离了加多宝的掌控。联合臭名昭著的公知大 V，派遣水军围攻“自干五”，甚至联合平台对抗政务大号，使得加多宝一步步卷入舆论漩涡。特别值得一提的是，在“@共青团中央”的带队下，全国青年网络力量在这一舆论仗中首次显示出了强大的团队作战能力。

10．“文登事件”，青少年互联网意识形态话语权争夺愈发激烈

2015 年 7 月 22 日，山东省文登市当地青年侯聚森在校园门口遭到 4 名外地青年的围殴，当地警方的处理结果受到网民的质疑，并引起高层的关注。“7·22 文登事件”对中国互联网意识形态领域的影响，将会随着时间的推移不断强化和加深，该事件的多焦点性令其舆情发酵呈现出波澜状的跌宕起伏。极端主义思潮在中国青年群体和学生群体的无序蔓延，使

得贴吧这个边缘舆论平台浮出水面，资本力量对舆论平台的掌控，更使得青少年意识形态领域呈现出巨大的危机。

上述十大网络舆情事件，是贯穿 2015 全年具有特别意义的经典舆情事件。除此之外，任志强炮轰团中央、锋锐“死磕律师”被捕、公知浦志强案、梁柱教授被网络围攻等年度网络舆情也对中国舆论场的未来变革，形成深远的影响。

1.2 网络舆论空间治理

截至 2017 年 6 月，我国的网民已超过 7.5 亿人，很多网民将互联网视为了解社情民意、揭露社会弊端、开展社会监督的窗口。2009 年以来，南京“天价香烟”事件、河南民工“开胸验肺”事件、云南晋宁县“躲猫猫”事件等热点事件，均由网络舆论率先关注，继而引发媒体报道。中国社科院发布的《蓝皮书》透露，在 2009 年 77 件影响力较大的社会热点事件中，由网络爆料而引发公众关注的有 23 件，约占全部事件的 30%。可见，互联网已成为新闻舆论监督的重要平台，特别是以开放性、匿名性及互动性为特点的微博、网络论坛等成为网络舆情的主要来源。

在大多数情况下，网络舆情易受人为因素影响，特别当话语权、传播权、定性权等舆论核心资源被社会媒体、网络大 V、舆情智库等舆论场金字塔顶端力量所掌控，网络舆情被人为炒作、扭转、掩埋的概率明显加大，衍生出来的舆论暴戾、网络谣言、谩骂攻击、虚假广告等，对网络舆情的发展和次生舆情的演变，乃至对社会的稳定、经济的增长、国家的发展都会产生重大的影响。

针对网络舆论空间中的种种乱象，需要通过治理与引导双管齐下，抑制网络舆论负能量，正确引导网络舆论走向。

1.2.1 网络炒作问题

随着网络舆论对社会和公众影响的不断增大，出现了以网络炒作为营生的网络公关公司、网络推手、网络水军等，网络公关公司受客户的委托，在网上炒作某个话题或人物来达到宣传、推销或者诋毁他人或产品的目的，为此雇佣了大量的网络推手、网络水军，在网络推手的组织下，网络水军以各种手法和名目在互联网的各大网络论坛上短时期内大量地发帖和回帖，炮制网络热点事件，捧红各色人物，营造虚假民意，形成虚假的网络舆情。例如，在央视感动中国 2010 年度人物评选中就遭遇网络水军的密集刷票，引起社会各界高度关注；通过网络炒作，使“奥巴马女郎”、“兽兽门”、“阎凤娇裸照门”、“凤姐”、“犀利哥”等原本无名人物在一夜之间名扬网络；在网络上被传得沸沸扬扬的“王老吉”添加门、“360”曝黑门、“康师傅”水源门、“伊利”牛奶门等热点事件都是通过网络炒家人为炒作出来的。

在网络炒作活动中，通常包括三类主体：客户、网络公关公司和网络水军。网络公关公司是客户与网络水军之间的中介，负责联系客户，得到任务，收取酬金，同时也负责招募、管理网络水军，发放任务和酬金等。其业务流程为：网络公关公司收到客户委托后，进行任务筹划和分工，将任务下发给网络推手（也称为水军头目），网络推手组织网络水军完成任务，并负责任务审核和酬金发放等。网络水军赚钱的模式为：领取新任务、完成任务、汇报任务、等待审核、审核通过、结算报酬。这样，网络公关公司、网络推手、网络水军就形成了灰色利益链，他们在实现客户目标的同时也获得自身利益。据公安部门调查，当前国内一些大的网络论坛，有 50%左右的帖子是人为炒作推出来的。所谓“热门帖”、“精华帖”等，很少是网民自发点击、回帖形成的，背后几乎都有网络炒家在积极推动，都是由网络水军实施的，这种虚假的网络舆情也称为网络灌水现象。

网络水军及其网络灌水问题具有很大的危害性，在网络舆情中存在歪曲失真信息泛滥、网民群体极化倾向严重、境内外不法分子恶意操纵、国外敌对势力渗透性入侵等隐患，产生错误的舆论导向，危及政府的公信力，引发社会群体性事件等问题。

对于网络炒作行为所产生的负面影响，已引起国家互联网管理部门的关注和重视，央视等主流新闻媒体多次对网络水军及网络炒作问题进行采访报道和深度分析；国家互联网管理部门制定了加强互联网管理的有关规定，并依法惩戒了利用互联网进行造谣惑众、恶意炒作的不法网民。

1.2.2　网络大 V 问题

所谓网络大 V 是指在新浪、腾讯、网易等微博平台上获得个人认证，拥有众多粉丝的微博用户，粉丝数量通常达到 50 万以上。“V”是指贵宾账户（VIP），账户会在名字后面显示一个 V 字符，表示是经过微博实名认证的高级账户，网民将这种微博用户称为“大 V”。

网络大 V 大多是有一定知名度的学者和名人，拥有大批的粉丝。在新浪和腾讯微博中，拥有 10 万以上粉丝的大 V 超过 1.9 万个、百万以上粉丝的大 V 超过 3300 个、千万以上粉丝的大 V 超过 200 个。网络大 V 们相当于意见领袖，其影响力不容小觑，往往成为爆料者的求助对象，他们的一次转发就会使得一条微博迅速火起来，成为网络热点话题，引导着网络舆论走向。

随着微博的兴起，网络水军把造谣和传谣的主要阵地从网络论坛转至微博，形成一条由网络推手、网络水军和网络大 V 组成的网络谣言制造和传播的灰色利益链。网络造谣者的商业模式往往是 “先赚名，再赚钱”，造谣本身未必产生收益，但通过造谣，吸引粉丝眼球，这些粉丝将成为日后商业炒作的基础。以名叫“立二拆四”的网络大 V 为例，通过这种商业炒作，其公司经营最好的时候年收入将近千万元，纯利润有数百万元之多。在造谣传谣的利益链中，网络推手、网络水军的造势再加上网络大 V 的转发影响力，才真正形成完整的将影响力变成金钱的利益链。

中央高度重视互联网管理和舆论治理，在《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》中指出，“牢牢把握正确舆论导向，健全社会舆情引导机制，传播正能量。加强网上思想文化阵地建设，实施网络内容建设工程，发展积极向上的网络文化，净化网络环境。”

近年来，国家互联网信息办公室陆续发布了《即时通信工具公众信息服务发展管理暂行规定（简称微信十条）》、《互联网新闻信息服务单位约谈工作规定（简称约谈十条）》等有关互联网管理规定，并落实到具体行动上。约谈、点名批评、惩戒了一些违反规定的网络大 V，如“立二拆四”、“秦火火”等。个别网络大 V 在突发重大公共事件中发表不当言论，或者以“求辟谣”、“求证”等方式故意扩散谣言，导致不明真相的网民跟风，损害了网络媒体公信力，也扰乱了网络正常传播秩序。

国家工商管理总局也发布了《互联网广告管理暂行办法》，指出网络大 V 在微博上转发广告帖，如果广告违法，网络大 V 要承担相应违法责任。

开展打击网络谣言等专项行动，使不法网络大 V 和网络谣言遭到压制性打击，一些被称为“推墙派”、“凿船派”、“体制婊”的网络负能量群体被有效遏制，网络空间雾霾渐散、晴空初现，网络舆论空间逐步呈现风清气正之势。

1.2.3 政务微博作用

政务微博在网络舆情引导上发挥着重要作用，在突发社会公共事件中，如果政府部门及时通过政务微博发布事件真相，与网民交流沟通，则能够有效抑制谣言传播，避免舆论风暴。这一点已经在多次的突发社会公共事件处置中得到充分的证明。

所谓政务微博是指代表政府机构和官员的、因公共事务而设的微博，主要是用于收集意见、倾听民意、发布信息、服务大众的官方网络互动平台，其目的主要在于通过与公众的良性互动，搭建一个社会化参政、议政、问政的网络交流模式与平台。

政务微博最早始于 2009 年下半年，湖南桃源县官方微博“桃源网”出炉，成为中国最早开通微博的政府部门。紧接着云南省委宣传部的官方微博“微博云南”面世，随后以“平安肇庆”、“平安北京”为代表的全国各地的公安微博，以及各级党政机关领导的微博如雨后春笋般开通。截至 2011 年 12 月，在新浪网、腾讯网、人民网、新华网四家微博客网站上认证的政务微博客总数为 50 561 个，其中党政机构微博客 32 358 个，党政干部微博客 18 203 个。在新浪网认证的党政机构微博客 12103 个，党政干部微博客 10 652 个；在腾讯网认证的党政机构微博客 13 911 个，党政干部微博客 6 748 个；在人民网认证的党政机构微博客 2 401 个，党政干部微博客 71 个；在新华网认证的党政机构微博客 3 943 个，党政干部微博客 732 个。

政务微博关注的对象分为两类：一类是关注其他省市自治区的相同部门的微博，及时获取同行的信息，充分借鉴经验、做法为自己所用，来改进本部门的工作；二是关注与自己所分管业务和行业相关领域的意见领袖（即网络大 V）的微博，这些意见领袖的微博都拥有

大量粉丝，具有深厚的相应领域专业知识和丰富的实践经验，一部分是本领域的专家学者，其言语和行为对其粉丝具有十分巨大的影响。政务微博只有充分关注自己分管领域的意见领袖的微博，才能及时知晓行业动态、存在的问题，甚至可以从其中找到解决问题的思路、解决问题的方法和改进本职工作的方式方法等。

开设政务微博的重要意义在于：

（1）听取群众呼声更便捷。微博便利了人们的沟通交流，通过微博，已经实现了手机终端和网络的互动，而由于手机终端已经相当普及，所以通过微博就能够非常便捷快速地交流。同时，微博简明扼要、直奔主题的文风也与简洁的会风高度契合。通过微博，网民可以不受时空限制地表达自己的意见和建议，这保障了人们的话语权，在民情民意的表达和收集方面具有明显的优势。微博的流行，调动了百姓参与公共事务、共商国是的积极性。

（2）开辟了政府发布信息的新通道。微博改变了网民的表达方式，改变了媒体的生态环境。通过发展政务微博，不仅有益于政府政策的公开和透明，还开辟了一条政府处理紧急事件的信息公开通道，同时也是对政府行政能力的考验。生活在网络信息时代，政府通过微博可以及时提供准确客观的重要信息。例如，在昆明市螺蛳湾商户聚集事件中，当天在网络上并没有出现以往群体性事件中曾经有过的种种谣言，也没有形成铺天盖地的舆论风暴。这都要归功于“微博云南”及时发布的 111 个字的信息。这使得政府改变了先处置、后发布的老套路，边做边说，变被动为主动，从起点就跑在了流言前面。

目前，政务微博在建设和使用上还存在不少的问题，例如：

（1）政务微博缺乏运行机制。政务微博运行处于自发状态、没有法律、纪律对其运行加以约束和限制。其主要表现一是信息发布不及时，信息发布数量不均。二是对用户提出的评论、意见和建议不能及时有效地回复和解决。三是互动功能不足，没有充分利用微博的各种交互工具来完成应有的功能。俗话说没有规矩无以成方圆，政务微博没有明确的规章制度和评价机制，就无法向广大用户提供及时、可靠的服务，同时也不利于政务微博的健康发展。

（2）政务微博缺少必要的营销手段。由于缺少营销理念和手段，使得政务微博的粉丝数和关注对象不足、发帖数量不足。发帖数量包括原创发帖、转发发帖和评价发帖等。发帖数量不足，将会直接导致内容更新不及时，信息量不足，从而造成已有粉丝自然流失。客观地说，一般情况和条件下，政务微博粉丝数的多少不是至关重要的，政务微博不能以追求粉丝数为终极目的。然而粉丝数越多，影响力越大也是不争的事实。当突发事件发生时，政府需要在第一时间播报事态发展情况，粉丝数的多少至关重要。粉丝数越多，政府的声音传达越迅速，有利于争取主动权，是防范流言和谣言的最好办法。

（3）政务微博的功能和定位不明确。由于政务微博发展时间不长，对政务微博属性并没有形成统一界定，特别是党政机构微博和党政干部微博之间的定位和功能不清，各微博客网站对政务微博的界定范围也存在差异。目前政务微博的主要功能还是信息发布和互动，实

际应用比较少，公众对政务微博的认识也还处于模糊状态，政务微博的政府机构和公务人员很多还处于自发状态。

（4）部分党政干部对微博的认识不全面。部分党政干部对信息化发展趋势，以及信息化发展对执政能力的影响缺乏全面、正确、深刻的认识，有的重视不够、认识不深；有的反应过度、防范过多；还有不少官员认为微博可能迅速传播自己的负面信息，不敢触及微博问政；部分党政干部对微博应用持警惕态度，对微博这一新工具存在恐惧和抵触心理。

（5）部分政务微博存在形式化、空心化、名利化现象。一些党政机构和干部虽然开通了政务微博，却不见经常更新或缺乏实质性内容；个别政府官员和政府机构开通微博后就不闻不问，成了“空壳微博”；还有一些政府官员和机构为了某种“形象名利”开设政务微博，摆花架子，缺乏实质性内容。多数政务微博仅仅作为单向信息发布工具，而没有充分利用微博客的互动特性，“说”的多，“听”的少，缺乏互动。

（6）缺乏统一标志，认证、评估监管和机制保障不足。政务微博的命名不规范，政府机构或公务人员微博的命名随意性较大，导致公众难以辨别真伪，这必然给政务微博的使用带来隐患，有被冒用、盗用制造混乱和谣言的潜在风险。政务微博作为政府官方微博，必须保证所发布和回应的信息准确、有效、及时，而要做到这一点，就需要一个由众多部门协同配合所构成的组织体系和运行机制予以支撑和保障。另外，对于如何发布信息、发布什么内容、按照怎样的流程处置，以及如何答复网民、引导舆论等都缺乏制度性规范。

因此，政务微博需要从以下几个方面进行改进，使政务微博在网络舆论引导上发挥更大的作用。

（1）明确目标，准确定位政务微博的功能。政府的执政理念和管理方法需要适应互联网时代的新要求，微博客特有的及时、互动、开放的信息传播特点形成了“微博问政”式的官民沟通新路径。政务微博作为一种新的自媒体交流工具，在促进民众信息分享、平等对话、参与社会管理等方面发挥了积极的作用。党政部门和公务人员要增强服务意识，正确认识政务微博的本质，按照社会管理创新、政府信息公开、倾听民众呼声、树立政府形象、有序参政议政等内容确定自身政务微博的功能，促进政务微博的健康发展。

（2）实施集群化整合，多渠道扩大影响。要重视政务微博的集群作用，整合区域和部门资源，通过粉丝收听、转发、评论等微博功能实现信息共享。管理和职能部门要加强统筹协调，充分发挥政务微博为民服务功能，同时加强与所在地媒体的互动。

（3）建立科学管理机制，促进良性发展。要建立政务微博开设、运营、管理的规章制度。通过建立信息采集制度，及时将有价值的信息政府对舆情的研判和把握能力；从舆情信息综合发现与挖掘、跨媒体与多通道内容的关联分析、舆情安全态势推演等整理报送职能机构和决策者，形成民意直通车，减少中间环节，提高行政效率；通过建立舆情评判机制，加强网络舆情掌控能力；通过健全工作机制，明确政务微博的响应时间、处置流程、改进办法，落实责任部门和团队，形成一整套与之相适应的管理机制。

（4）加强与政府网站的有机结合。政务微博应与政府网站有机结合起来，完善协同配合功能，对信息公开、舆论引导、政民互动、政策宣传、引导动员、社会监督等内容进行梳理，设计政府内部响应、处理、回应机制和工作流程，与政府网站形成整体性效能统一的运行管理体系。

（5）开展微博客应用绩效评估，促进工作持续改进。政府的执政理念和管理方法需要适应互联网时代的新要求，微博客特有的及时、互动、开放的信息传播特性，通过设计科学合理的评估指标体系及切实可行的政制，促进政务微博健康可持续发展。

（6）提升党政干部信息素养，提高政务微博运营质量。党政干部应当了解网络运行规律，提高网上信息的甄别能力、网上舆情的研判能力和网络舆论的引导能力。要善于利用网络搜集信息，及时把握社情民意，以适应互联网环境下的政府管理创新和服务型政府建设。

1.3　网络舆情传播平台

网络舆情主要通过网络新闻、微博、网络论坛等平台传播的，尤其是以开放性、匿名性及互动性为特点的微博、网络论坛等成为网络舆论的主要来源。微博和网络论坛是两种不同的信息传播平台，其网络信息传播机制也是不同的。

1.3.1　微博网络

微博网站是一种集成化、开放式的社交服务平台，用户通过 140 字以内的微博发布信息，实现即时分享。同时，用户还可以选择关注其他用户，关注其信息，而且也可以被其他用户相互连接，交流信息。可见，微博平台具有社交网络和媒体网络的双重特性。

Twitter 网站是世界上最早出现的微博网站，最初的服务比较简单，主要提供向好友的手机发送文本信息的服务，现在已发展成一个集社交网络和微博为一体的综合社交服务平台。后来，国内的主要门户网站也相继开设了微博网站，如新浪微博、腾讯微博、搜狐微博等，其中新浪微博是国内最大的微博平台，其注册用户数超过 5 亿人，日活跃用户数达到 4 620 多万人，微博用户数量迅猛地增长。尽管近几年受到微信等即时通信工具的冲击，但微博的网民数量仍然是比较庞大的。

微博作为新兴的社交媒体，越来越受到重视。在国外，很多的政治人物、政府部门、新闻机构等都开通了 Twitter 账号，作为与民众沟通交流、获取信息的手段。在国内，一些政府部门陆续开通了政务微博，实时发布消息，与民众互动；CCTV、人民日报、新华通讯社等主流媒体也都在新浪等微博平台上开通了官方微博，作为新闻发布、了解民意、监测舆情的主要渠道。根据人民网网络舆情监测室的 2015 年突发公共事件舆情分析报告，在各种突发公共事件的舆论关注度中，微博通常达到万条以上，而网络新闻、论坛帖子通常不超过千条。可见，微博已经成为网络舆情的主要来源地。

微博转发是微博网络的主要信息传播机制，用户可以将关注者发布的微博转发到自身平台上，然后分享给粉丝。通过这种信息传播机制，使得一条微博能够在更大范围内传播和分享。可见，用户转发行为是推动微博信息传播的重要因素。

由于微博网络是一种社交网络，用户转发行为与用户的社会纽带关系密切相关。在微博网络中，用户之间存在三种社会纽带关系：强连接、弱连接及权威连接，不同的社会纽带关系对用户转发行为的影响也不同。因此，在分析用户转发行为时，首先需要识别用户之间的社会纽带关系。

强连接关系通常表示用户彼此之间具有高度的互动，在某些存在的互动关系形态上较为亲密。因此，通过强连接传播的信息通常是重复的，容易自成一个封闭的系统。网络内的成员由于具有相似的态度、高度的互动频率通常会强化原本认知的观点，而降低了与其他观点的融合，强连接网络并不是一个可以提供创新机会的渠道。

相对于强连接关系，弱连接关系则能够在不同的团体间传递非重复性的信息，使得网络中的成员能够增加修正原先观点的机会。事实上，在信息扩散传播方面，弱连接起着同样的作用。一个人的亲朋好友圈子里的人可能相互认识，在这样的圈子中，他人提供的交流信息总是冗余的。例如，从这个朋友或亲戚听到的信息，可能早已经从另一个朋友那里听说了，而他们之间也都相互交谈过此话题。日常生活中不乏这样的事例。

权威连接完全不同于强连接和弱连接，主要表现在用户之间的非对称性，非对称性包括两个方面，一是用户影响力的非对称，例如网络大 V 的影响力比一般用户大很多；二是信息传播的非对称性，例如网络大 V 的微博很容易被一般用户转发，而一般用户的微博很难被网络大 V 转发。在权威连接关系中，信息传播方向一般由权威高的用户到权威低的用户。在社会科学中，这种现象称为服从权威。

在社交网络中，信息传播存在两个重要进程：同化与社会影响。同化是指信息在网络传播过程中容易导致用户与自己观点、价值观相似的用户建立连接关系，最终使社交网络的结构发生变化。社会影响则是指信息在网络传播过程中导致相邻用户的观点、价值观等属性逐渐趋于一致，最终使两个用户具有相似性。因此，在信息传播过程中，不同社会纽带关系的用户将受到同化和社会影响的影响，最终导致了网络结构和用户属性的变化，也就是说，不同的社会纽带关系将有不同的网络结构和用户属性。反过来说，通过提取网络结构和用户属性等特征，就能够识别出用户之间的社会纽带关系。网络结构特征可以从微博用户关注图得到，其中包括权威比率、微网络结构；而用户属性则可以从用户的个人资料提取出来。

因此，在研究用户转发行为时，需要根据权威比率、微网络结构、地理距离以及性别等特征识别出潜在的社会纽带关系，分析各个特征间的相关性以及用户转发行为的内在动力，为研究微博舆情形成机理，正确引导微博舆论提供科学依据。

1.3.2　网络论坛

网络论坛是一种开放性、匿名性及互动性的信息交流平台，网络论坛类型多种多样，如综合性论坛、专题性论坛等，涉及内容涵盖了社会生活的方方面面，每个网民都可以找到自己感兴趣或者需要了解的专题论坛，促进了网民之间的交流，增强了网民的互动性。

网络论坛属于传统的网络信息交流平台，随着社交网络、微博网络等新型网络信息交流平台的广泛应用，网络论坛的用户数量有所下降，尽管其网民数和使用率不如微博、社交网络高，但网络论坛所具有的多元化、开放性、匿名性及互动性，仍然是广大网民发表言论、获取信息的重要网络平台，用户数量还是比较庞大的。

网络论坛的最大特点是开放性、匿名性和隐匿性，用户不需要实名制注册，可以随意注册多个不同的用户名而不用泄露自己的真实身份；用户只要登录网络论坛，就可以随意发布或回复信息；用户在登录的情况下可以浏览网络论坛中的全部信息，而不受好友关系限制，甚至在不登录的情况下也可以浏览网站大量内容。因此，网络论坛不仅成为网络舆情的主要来源地，也是网络水军进行网络炒作、造谣传谣的主要平台。

在网络论坛中，网民就某个主题通过发帖、观看和回帖进行信息交流和互动，在信息交流过程中，某些话题的帖子受到网民的高度关注，点击量和回帖数非常大, 形成较大的影响力，这种帖子称为热帖，热帖在观点传播和舆论形成过程中起到重要的推动作用。可见，网民通过发帖和回帖发表意见，参与观点传播和舆论形成，成为网络舆情的主要来源。

在网络舆论形成过程中，意见领袖起到了积极的推动作用。统计数据显示，网络中的大部分用户不经常参与信息的制造与传播，他们做出的决定往往跟随意见领袖。通过意见领袖发表引导性意见来影响所在网络用户而非直接说服他们，可以有效地触发整个网络舆论的影响力，能够有效地推动信息的传播，提高广告效应。同时，网络论坛也是一把双刃剑，它所具有的开放性和匿名性等特点，容易被别有用心的组织和人员所利用，传播虚假消息和谣言，对人们的社会生活和意识形态造成负面的影响。

目前，对虚假网络舆情的界定还缺乏统一的标准和共识，通常从网络舆情所表现出的外在特性来识别虚假网络舆情。由于虚假网络舆情是由网络水军操纵而形成的网络舆情，在时间特性、空间特性、主题特性以及情感特性等方面与网民自然发帖、回帖而形成的网络舆情具有明显的差异，通过分析网络舆情所表现出的外在特性，能够识别出虚假的网络舆情。

（1）时间突发性强：虚假网络舆情通常具有明确的话题指向。在网络论坛中，话题对应于网络论坛中的帖子。网络水军接受雇主（网络公关公司或网络推手）指派的任务后，会按照雇主意旨编撰网帖，然后为了形成集束效应以左右网络舆论、吸引网民关注，在短时间内大量地发帖和回帖，舆情信息像病毒一样在网络中大肆蔓延，形成虚假网络舆情。

（2）主题相似性高：对于虚假网络舆情，其传达的舆情信息具有明确的主题倾向性，呈现一边倒的状况。同时，由于是网络水军幕后推动产生的结果，为了博取网民信

任，在评论内容的语气和书写习惯上也非常相似。一般来说，网帖内容或捧或骂、篇幅简短、态度鲜明。

（3）空间相似性大：由热点事件引发的网络舆情绝大部分都集中在国内有影响力的网络论坛上，参与有关事件话题讨论的网民分布在全国各地。即使普通的热点事件，相关文章的 IP 地址出处也比较分散。而由网络水军推动的热点事件因推动者所处地域比较集中，即使发帖的用户 ID（即用户标识符，俗称“马甲”）不同，但其 IP 地址一般集中在一个相对较小的 IP 地址空间里。

（4）新注册用户 ID 比例高：网络水军为了隐藏自己的身份，营造出舆论来自于现实世界网民声音的假象，同时为了突破一些网络论坛对用户“单日发帖数量”的限制，通常注册多个用户 ID 来发表话题评论，这些用户 ID 基本是为推动某一热点事件而专门注册的，注册时间较短。而由网民自然发帖、回帖形成的网络舆情，其用户 ID 的注册时间分布是比较均匀的。

（5）用户 ID 离散度高：网络水军为了提高所炮制事件的热度，吸引广大网民加入讨论，同一水军会反复不断地以不同的“马甲”发表相似的看法，但是对应的 IP 地址是相同的，具有较高的用户 ID 离散度。对于可以得到用户 ID 的网络论坛，可以通过统计某一热点话题的用户 ID 离散度来识别是否为虚假网络舆情。

1.4 网络舆情分析技术

随着网络舆论影响力的增强，网络舆情已经成为各级政府了解社情民意、改进工作作风、提高执政能力的重要窗口。近年来，国家大力推进网络舆情监控体系建设，各级宣传主管部门以及主流新闻媒体等大多设立了网络舆情监测机构，加强对网络舆情监测和引导。

在网络舆情监测中，面对海量的网络信息，必须借助于信息技术来实现网络舆情监测的自动化和智能化。于是，各种网络舆情监测系统便应运而生了。

1.4.1 网络舆情监测系统

网络舆情监测系统的主要功能是实现网络信息的自动采集和网络舆情的在线监测与分析。尽管不同的网络舆情监测系统产品存在一定的差异，但在系统架构和核心技术上大同小异。通常，一个网络舆情监测系统可以按照层次化结构来构建，主要分为数据采集处理、舆情分析引擎和舆情分析服务等三个层次，如图 1-2 所示。

（1）数据采集处理层：主要提供网络数据采集和预处理功能，网络数据监测和采集的对象主要是主流的网络新闻、微博、网络论坛、网络博客等网站的文本信息，对于采集到的网络数据，首先需要进行初步的数据过滤、去重等预处理，经过数据格式转换及元数据标引后，存入数据库待进一步处理。

（2）舆情分析引擎层：主要提供话题检测、话题跟踪、倾向性分析、自动摘要以及中文分词等功能，舆情分析引擎是网络舆情监测系统的核心功能，主要完成热点话题的检测、跟踪以及情感倾向性分析，并且对各类热点话题及倾向性进行自动摘要，分析结果存入数据库，以便为用户提供各种舆情分析服务。舆情分析引擎的核心技术是文本聚类、文本分类、情感分析中所采用的模型与算法，直接关系到系统的性能高低。不同的网络舆情监测系统所采用的模型与算法可能有所不同，系统性能也会不同。

（3）舆情分析服务层：主要提供突发事件分析、舆情预警报警、舆情趋势分析、舆情统计报告以及舆情查询检索等各种舆情分析服务，以方便用户使用。

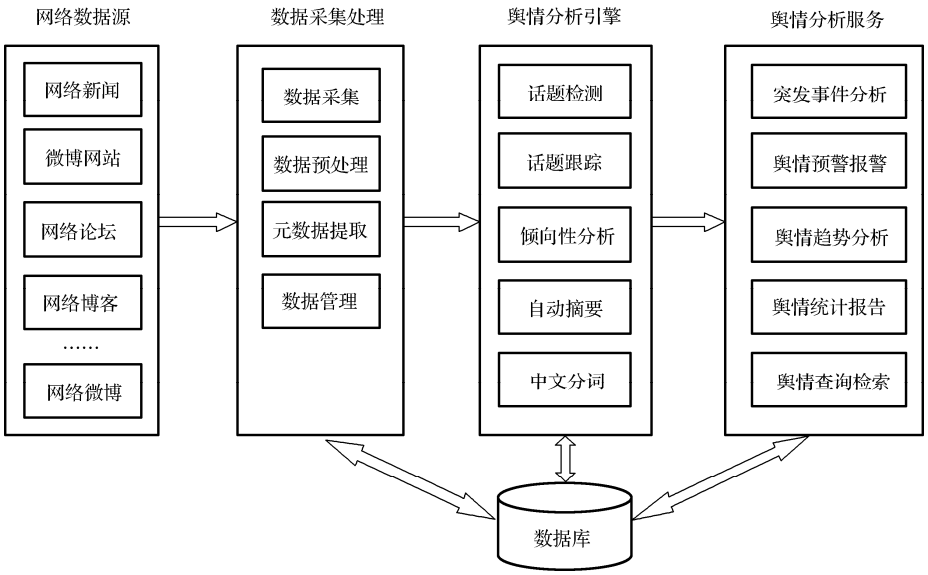


图 1-2　网络舆情监测系统层次结构

网络舆情监测系统通常采用基于客户/服务器的系统结构，系统分为客户机和服务器两个部分，服务器是整个系统的核心，系统核心软件运行在服务器上，提供数据采集、舆情分析、舆情服务等核心功能，并为用户舆情分析服务。客户机为基于浏览器的用户界面，用户使用浏览器登录到服务器的用户界面上，按照用户界面提供的功能菜单，获得系统提供的各种舆情分析服务。

可见，一个网络舆情监测系统的核心技术是网络舆情分析技术，主要涉及网络信息采集技术、话题检测与跟踪技术、文本情感分析技术等。

1.4.2　网络信息采集技术

网络舆情的数据来源是互联网中各种网络媒体、信息交流平台发布的网页信息，其中

包括网络新闻、微博、论坛帖子等，尤其是微博、网络论坛等交互式网络平台，成为网络舆论的主要来源地。因此，在网络舆情分析中，首先需要使用网络信息采集工具自动搜索和采集主要网络媒体网站或平台发布的信息，为网络舆情分析提供数据资源，网络舆情分析的效果在很大程度上取决于网络信息搜索和采集的质量。

网络信息采集技术涉及搜索引擎、网络蜘蛛、网页搜索算法、网页相关性分析等相关技术。

1．搜索引擎

搜索引擎是互联网中最常用的信息搜索工具，著名的搜索引擎有 Google、百度等。搜索引擎自动搜集互联网中的网页信息，经过整理、组织、加工和处理后，建立管理和存储这些信息的索引库，并提供基于索引的信息检索服务。当用户发出搜索请求时，搜索引擎便根据用户提交的查询条件，从索引库中迅速查找出用户所需的信息，并返回给用户。

搜索引擎通常由网络蜘蛛、索引器、检索器和用户接口等部分组成，网络蜘蛛在互联网中不断地搜索，发现和采集新的网页信息，然后将网页信息存入网页库，由索引器建立索引；索引器将分析网络蜘蛛所采集的信息，从中抽取出索引项，建立用于检索页面的索引表；检索器将根据用户的查询请求和条件，从索引库中快速检索出网页，并通过网页相似性分析和评价，对输出的结果进行排序；用户接口为用户提供一个输入查询请求和显示查询结果的用户界面。

根据信息搜索范围，搜索引擎可以分为通用搜索引擎和主题搜索引擎，通用搜索引擎的搜索范围比较广泛，但搜索出的无关信息较多；主题搜索引擎是针对特定主题的信息搜索，具有“专、精、深”特点。按照信息搜索方式和服务提供方式，搜索引擎可以分为目录搜索引擎、机器人搜索引擎、元搜索引擎等类型，其中，机器人搜索引擎也称为网络蜘蛛或网络爬虫，也是最常用的搜索引擎。

2．网络蜘蛛

网络蜘蛛也称网络爬虫，是搜索引擎的核心部件。在搜索引擎中，网络蜘蛛主要负责搜集网页、图片和文档等信息。其搜集过程是从给定的起始 URL 开始，沿着网页中的链接，按照一定的搜索策略进行遍历搜索，下载相应的网页，解析出网页中的超链接 URL，将那些未访问过的 URL 加入到待搜索队列中，然后再搜索其他链接指向的网页，循环往复。整个过程如同一个蜘蛛在蜘蛛网（Web）上爬行。

网络蜘蛛在搜集网页时采用两种搜索策略：深度优先搜索策略和广度优先搜索策略。深度优先搜索策略是指网络蜘蛛从起始网页开始，一个链接一个链接地搜索下去，处理完这条路径之后再转入下一个起始网页，继续跟踪链接，直到遍历所有的网页及链接，搜索过程结束；广度优先搜索策略是指网络蜘蛛从起始网页开始，首先搜索完一个网页中所有的链接，然后再继续搜索下一层，直到底层为止。广度优先搜索策略通常是网络蜘蛛的最佳搜索策略，不仅容易实现，并且还能够实现并行处理，提高其搜索速度。

网络蜘蛛同样也可分为通用蜘蛛和主题蜘蛛，与通用蜘蛛相比，主题蜘蛛更加专业化和可定制化。通用蜘蛛的目标是尽可能多地采集网页信息，而不太关注网页采集的顺序和被采集网页的主题。主题蜘蛛能够定向性地采集与主题相关的网页，忽略无关的网页，并且还可以根据主题相似度值进行优先采集。

3．网页搜索算法

网页搜索算法是网络蜘蛛的核心，它采用一定的搜索策略来搜集网页资源，尽可能多地搜集与主题相关的网页，同时也要尽可能少地搜集无关的网页，以保证网页的搜集质量。目前，人们提出了多种搜索策略，如基于链接结构评价的搜索策略、基于网页内容评价的搜索策略等。

基于链接结构评价的搜索策略是利用 Web 结构信息来指导搜索，并通过分析 Web 网页之间的相互引用关系来评价网页和链接的重要性。这种策略的基本思想来自于文献计量学的引文分析理论，将引文分析理论应用于 Web 环境时，主要采用基于链接结构的评价方法。采用这种策略的搜索算法有 PageRank 算法、HITS（Hyperlink-Induced Topic Search）算法等，这两种算法的共同点是利用网页之间的引用关系来确定链接的重要性，充分考虑了链接的结构特征，但也存在一些缺陷：一是忽略了网页与主题的相关性，在某些情况下，可能会出现搜索偏离主题的“主题漂移”问题；二是计算复杂度将随访问网页和链接数量的增长呈指数级增长。

基于网页内容评价的搜索策略是利用网页文本内容作为领域知识来指导搜索，并根据网页文本与主题之间相似度的高低来评价链接价值的高低。采用这种策略的搜索算法有 Fish Search 算法、Shark Search 算法等，Fish Search 算法是一种基于客户端的搜索算法，根据用户的种子站点和查询的关键词或短语，将包含查询字符串的页面看作与主题相关，计算该网页与主题的相似度，动态地维护待搜集 URL 队列。Shark Search 算法是对 Fish Search 算法的一种改进，主要改进了网页与查询信息相似度计算方法。

4．网页相似度计算

在主题蜘蛛中，需要对搜集的网页内容与查询的主题内容进行相似度计算，判别它们是否相关。因此需要采用适当的表示模型来描述文本，使之能够对网页内容和查询内容之间的相似度进行量化计算，准确地评估网页相关性。常用的表示模型是向量空间模型（VSM），该模型具有算法简单、计算复杂度低等特点，比较适合对网页文本内容进行实时处理。

在向量空间模型中，通过称为项的向量来表示用户的查询要求和文档信息，根据向量空间的相似度大小来排列查询结果。项也称为特征词，作为表示文档内容特征的基本语言单位，如字、词、词组或短语等。向量空间模型将查询词和文档按照特征词的维度分别进行向量化，然后通过适当的相似度度量方法计算出文档与查询词的相似度，优先检索那些与查询词相似度大的文档，并按照与查询词的相似度对检索出的文档进行排序。向量空间模型不仅

可以方便地产生有效的查询效果，而且还能提供相关文档的文摘，对查询结果进行分类，为用户提供准确定位所需的信息。

在网页相似度计算时，首先需要对一个句子进行分词处理，即按照词的含义对一个句子进行切分，将连续的字串或序列按照一定的规范重新组合成词序列，以便机器理解。汉语分词比英文要复杂，常用的汉语分词方法有正向最大匹配分词、逆向最大匹配分词和基于统计的词网格分词等。

综上所述，网络信息采集技术应用比较广泛，技术也比较成熟。由于网络舆情主要是通过网络新闻、微博、网络论坛等网络媒体进行传播的，这些网络媒体通常为动态网页，以松散的非结构化信息为主题，使得对动态网页的信息采集存在一定的困难，一些搜索引擎采取消极的规避策略来尽量避免过多地采集动态页面信息，这样会影响到信息采集的准确率和覆盖率。另外，新浪微博等微博平台出于保护用户隐私信息的目的，对微博信息的采集进行了限制，也影响到对微博舆情监测与分析效果。

1.4.3 话题检测与跟踪技术

话题检测与跟踪（TDT）的研究最初是由美国国防高级研究计划署（DARPA）发起的，旨在没有人工干预的情况下自动检索、判断和识别新闻数据流中的话题，通过每年举行的 TDT 测评会议，发表和展示 TDT 研究成果，并确定 TDT 研究方向和课题，以及 TDT 测评任务。TDT 测评会议共设立了 6 项测评任务，即：新事件检测（New Event Detection）、报道关系检测（Story Link Detection）、话题检测（Topic Detection）、话题跟踪（Topic Tracking）、自适应话题跟踪（Adaptive Topic Tracking）和层次话题检测（Hierarchical Topic Detection），其中话题检测与话题跟踪是核心问题。

TDT 技术的最初应用主要是新闻出版领域，用于新闻流的话题检测和事件跟踪。后来被扩展到互联网上，用于检测和跟踪以话题词为中心的互联网新闻热点话题以及流行词，因此成为网络舆情分析中的重要技术。

TDT 是从一篇文章的主题作为切入点，通过对文章主题的发现和跟踪，把各种分散的信息进行有效的汇集，并且组织成线索提供给用户进行查阅，厘清一个主题事件的来龙去脉，把握整个事件的整体和细节。例如，在网络舆情监测中，通过 TDT 技术对各种信息源的监测和分析，从中识别出针对某一突发事件的各种报道，并对事件的演化过程进行跟踪。TDT 技术还可以应用于证券市场分析等领域，用途比较广泛。

TDT 技术主要涉及报道和话题的表示模型、相似度计算、特征项权重计算、话题和报道间的相似度计算、文本分类与聚类的策略选择等相关技术。

1．表示模型

为了判断一个报道是否与某一话题相关，首先需要使用适当的模型来表示报道和话

题，以便对两者的相关性进行计算和比较。常用的表示模型有向量空间模型和语言模型。其中，语言模型是一种概率模型，语言模型的基本思想是对于在某一报道中出现的词，采用期望最大化（EM）等算法来分别估算该词在某个话题所有报道的概率分布和在整个语料库中的概率分布，可以得到某一报道讨论该话题的概率，这样就构成了一个词的生成模型。

在话题检测与跟踪中，人们提出了多种语言模型，如隐马尔可夫模型、指数语言模型、层次语言模型、语义模型等，其中效果较好的是 LDA（Latent Dirichlet Allocation）模型。

2．相似度计算

在 TDT 中，为了判断某个报道属于哪个主题，首先需要采用某种相似度度量方法来计算报道和主题之间的相似度，然后将相似度值和阈值进行比较，最后做出判断。相似度度量方法有很多种，TDT 中常用的相似度度量方法有内积、Dice 系数、Jaccard 系数、余弦系数以及欧几里得度量等。

3．特征项选取

在向量空间模型中，使用特征项来表示文本向量空间中的各个维度，因此特征项选取方法非常关键。直接使用分词和词频统计方法来得到特征项，可能造成向量空间维度比较大，给后续处理带来很大的计算开销，还会影响到分类和聚类算法的性能。因此，需要对文本向量做净化处理，在保证原文含义的基础上，找出最具代表性的文本特征项。这个问题归结为找到一种低维度的特征选择方法。最常用的特征选取方法是统计方法，这种方法比较精确，人为因素的干扰较少，尤其适合于文本自动分类挖掘。

基于统计的特征选取方法通过构造评估函数，对特征集合中的每个特征进行评估和打分，这样每个词语都获得一个评估值，又称为权值。然后将所有特征按权值大小排序，提取预定数目的最优特征作为提取结果的特征子集。这种方法关键是评估函数的性能，决定了文本特征提取的效果。这类算法主要有文档频率（DF）、信息增益（IG）、互信息（MI）、卡方检验（CHI）等，其中 CHI、IG 和 DF 的性能较好。

4．文本聚类

话题检测是一个文本聚类问题，其任务是将某个话题的所有报道自动归入一个话题类，它是在事先没有分类体系和训练语料的情况下对报道进行聚类分析，给出一个最佳的划分，而不需要预先对文档类别进行标注。

文本聚类是一种无监督的学习过程，不需要预先对文档进行手工标注类别，即不依赖于文档集合划分的先验知识，仅仅根据文档集合内部的文档对象彼此之间相似度关系并按照某种准则进行文档集合划分。文本聚类划分主要依据于这样的聚类假设：同类中的文档彼此之间的相似度较大，而不同的类之间的文档相似度较小。由于文本聚类分析不需要事先定义文档类别，对获取大规模多元数据集合的结构特征是有效的，能够发现数据之间所隐含的某些关系，因此在数据挖掘和知识发现领域中得到了广泛应用。

典型的文本聚类过程可以分为三个步骤：文本表示、聚类算法和效果评估。文本表示是指使用向量空间模型等文本表示模型，把文档表示成聚类算法能够处理的形式；聚类算法是指使用无监督学习算法对文档集合进行划分，文本聚类算法有很多种，常用的算法有层次方法、划分方法、基于密度的方法、基于网格的方法、基于模型的方法等；效果评估是指使用准确率、召回率、漏报率和误报率等测评指标来评价聚类的效果，也是对聚类算法性能的评价。

5．文本分类

话题跟踪是一个文本分类问题，其任务是判断某个报道是描述了一个新话题还是对某个旧话题的进一步跟踪报道。话题跟踪是一种特殊的文本分类过程，与传统的文本分类过程相比，话题跟踪中的文本分类是面向话题而不是面向概念更宽泛的主题，判断的依据更具体、粒度更细，处理的对象是动态的、随时间变化的报道流，而不是静态的文本集合。因此，在话题检测和跟踪中，不遗漏信息更为重要。

文本分类是一种有监督的学习过程，需要事先给定一个分类体系和一个标注好类别的文本集合，利用这些资源来构造一个分类器，将待分类文本归入不同的、预先定义的类别中，可以把这种分类过程称为文本归类。

文本分类过程可以分为手工分类和自动分类，手工分类首先由专家定义分类体系，然后由人工进行网页分类。这种方法需要大量的人力，现实中已经很少采用了。自动文本分类方法大致可以分为两类：知识工程方法和机器学习方法。两者相比，机器学习方法能够达到相似的精确度，并减少了大量的人工参与，成为文本分类的主流方法。

典型的文本分类过程可以分为三个步骤：文本表示、分类器构建和效果评估，其中文本表示和效果评估的方法与文本聚类相同，而分类器构建是文本分类中关键的环节，应当根据所要解决问题的特点来选择一个分类器。在选定构建方法之后，在训练集上为每个类别构建分类器，然后把分类器应用于测试集上，得到分类结果。在文本分类中使用的学习算法有多种，如 Rocchio 算法、k 最近邻居（KNN）、决策树、简单贝叶斯、神经网络、最大熵、支持向量机（SVM）等。其中，比较常用的是 Rocchio、KNN、决策树、SVM 等算法。

事实上，每种分类算法都有各自的长处和局限性，它们经常可以互为补充。实际应用和算法实验表明，在文本分类中，KNN 方法和多种方法的组合具有较好的性能。

1.4.4 文本情感分析技术

在网络舆情监测中，对于一个突发社会公共事件引发的网络舆情，网民所持有的情感倾向性往往是多元化的，包括正面或负面、赞扬或批评、支持或质疑、肯定或否定等。通过文本情感分析技术，能够自动识别出其情感倾向性，并给出分类统计结果，有助于及时采取应对措施。

文本情感分析技术主要研究如何对文本所表达的观点、情感、立场、态度等主观性信息进行自动分析，从海量文本中识别出人们对某一事件或政策等所持有的观点是褒义还是贬义，提高对文本情感分析的效率。文本情感分析技术涉及自然语言处理、计算语言学、人工智能、机器学习、信息检索、数据挖掘等多个研究领域，属于交叉性技术。

文本情感分析可以分为词语情感分析、句子情感分析、段落情感分析、文档情感分析等不同的层次。

词语情感分析的对象是在特定的句子中出现的词和短语。表达情感的词大多是名词、动词、副词和形容词，其情感倾向可以分为褒义、贬义和中性等三类，词语情感分析包括对词的情感极性、情感强度以及上下文模式等进行分析。在词语情感分析时，需要借助于标注有倾向性的情感词典，通常是面向领域应用来构建情感词典。在构建情感词典时，大多采用在已有的电子词典或词库上进行扩展的方式。例如，在知网（HowNet）的知识库上进行扩展。

句子情感分析的对象是在特定的上下文中出现的句子，其目的是通过分析句子中的各种主观性信息，判断该句子是主观句还是客观句。对于主观句，进一步提取出句子中的主观关系，实现对句子的情感倾向的判断，同时还要分析与情感倾向性相关的各个要素，如评价对象、情感极性、情感强度等。由于文本情感分析的对象是主观句。因此，主题句、主观句以及主观关系等识别和提取是句子情感分析的基础。

段落情感分析的对象是经过文本分割后的语义段而不是自然段落。由于语义段之间存在着语义联系，因此有助于对文本情感进行细化分析。在语义段情感分析时，以语义段中的句子为基本单元，通过计算句子情感值和语义段情感值，最终得到文本的全局情感值，实现对整个文本的情感分析。

文档情感分析的对象是一篇完整的文章，从整体上分析某个文章的情感倾向性。由于文档情感分析属于文本分类问题，通常采用机器学习方法，如朴素贝叶斯、最大熵、支持向量机等方法来解决文本情感分析问题，首先构建语料库，人工标注语料库中每个文本的情感倾向，并将语料库分为训练集和测试集，然后对模型进行训练和算法测试，对模型和算法的文档情感倾向识别能力进行评价。

在文本情感分析中，主要采用有监督的机器学习算法来识别文本中的评价对象及情感倾向。这种方法需要事先由人工标注语料库的情感倾向，作为训练样本，不同领域的训练样本也不同。然后构造一个分类器算法，经过自动训练后，对待分析文本的情感倾向进行分类识别。这种方法的优点是简单易行、识别准确率较高，整体效果较好。但是该方法依赖于人工标注的语料库，而人工标注语料库费时费力，并且缺乏标注标准，语料库标注格式也不统一。

另外，在文本情感分析中可以采用语言建模方法，它采用统计学和概率论方法对自然语言进行建模分析，发掘出自然语言中的规律和特性，解决自然语言信息处理中的特定问

题。语言建模技术已被广泛应用于语音识别、光学字符识别、手写字识别、机器翻译、文本分类以及文本检索等诸多领域，成为自然语言信息处理的主流技术之一。在基于语言建模的文本情感分析中，首先选择一种统计类语言模型作为基本语言模型，然后在标注有褒贬倾向的训练文本集上对情感模型进行估计。对于每一个测试文本，比较其语言模型与情感模型之间的相似度，如果与某个情感模型更为相似，则认为该文本的褒贬倾向与这个模型的褒贬倾向相一致，从而实现对文本情感倾向的识别。

由于文本情感分析技术将文本的情感倾向分为褒义和贬义两类，对于网络舆情监测中来说，还不够细致。在此基础上，还需要通过人工做进一步的统计分析。

第2章

网络信息采集技术

2.1 引言

网络舆情分析的对象是来源于互联网中各种信息交流平台发布的网页信息，因此网络舆情分析的首要条件是搜集互联网中网页信息。在搜集网络信息时，需要借助于专用的网络工具，如搜索引擎等，著名的搜索引擎有谷歌（Google）、百度（Baidu）等，也是网民最常用的网络信息搜索工具。

搜索引擎采用某种搜索策略在互联网上搜集网页信息，然后对信息进行提取、整理、组织和处理，建立索引数据库，为用户提供信息检索服务，起到信息导航的作用。搜索引擎的出现在很大程度上缓解了人们在互联网上查找信息的困难。经过多年的发展，搜索引擎的功能越来越强大，提供的服务也越来越丰富，成为广大网民不可缺少的网络工具。

网络舆情分析的数据来源是互联网中各种网络媒体、信息交流平台发布的网页信息，尤其是互动式信息交流平台或网站，如论坛、微博等，成为网络舆论的主要来源地。因此，在网络舆情分析中，首先需要使用网络信息采集工具自动搜集主要新闻网站、信息交流平台发布的信息，为网络舆情分析提供数据资源。网络舆情分析的效果在很大程度上取决于网络信息搜集的质量。

本章主要介绍与网络信息采集技术相关的搜索引擎、网络蜘蛛、网页搜索算法、相似度计算、主题蜘蛛组成等内容。

2.2 搜索引擎概念

2.2.1 通用搜索引擎

目前，在互联网上使用的搜索引擎有很多，如谷歌、百度等，这些搜索引擎主要关注的是广大用户的信息搜索需求，这类搜索引擎也称为通用搜索引擎。通用搜索引擎将自动搜索互联网中各种信息，经过整理、组织、加工和处理后，通过建立索引数据库来管理和存储

这些信息，并提供基于索引的信息检索服务。当用户发出搜索请求时，搜索引擎根据用户提交的查询条件，从索引数据库中快速检索出用户所需的网页信息，并返回给用户。

1．通用搜索引擎分类

按照信息搜索方式和服务提供方式的不同，搜索引擎可以分为如下三大类。

（1）目录搜索引擎：以人工方式或半自动方式搜集信息，由编辑人员查看信息后，人工生成信息摘要，并将信息放置在事先确定的分类框架中。信息通常面向网站提供目录浏览服务和直接检索服务。这类搜索引擎因加入了人的智能，所以信息定位准确、导航质量高，缺点是需要人工介入、维护量大、信息量较少、信息更新不及时等。这类搜索引擎的代表有雅虎、LookSmart 等。

（2）机器人搜索引擎：由一个称为网络蜘蛛或网络爬虫的机器人程序以某种搜索策略自动地在互联网中搜索信息，并为搜索到的信息建立索引数据库，为用户提供信息检索服务。这类搜索引擎的优点是信息量大、更新及时、无须人工干预。缺点是返回信息过多，有很多无关的信息，用户必须从结果中进行筛选。这类搜索引擎的代表有谷歌、百度等。

（3）元搜索引擎：元搜索引擎是将用户的查询请求同时递交给多个搜索引擎，将返回的结果进行重复信息排除、重新排序等处理后，作为最终的结果返回给用户。这类搜索引擎的优点是返回结果的信息量更大、更全。缺点是不能充分利用所使用搜索引擎的功能，用户需要做更多的筛选。这类搜索引擎的代表有 Web Crawler、Info Market 等。

除了上述的主流搜索引擎外，还有一些门户网站也提供信息搜索及其查找服务。

2．通用搜索引擎结构

通用搜索引擎通常由网络蜘蛛（Spider）、索引器、检索器和用户接口等 4 个部分组成，其系统结构如图 2-1 所示。

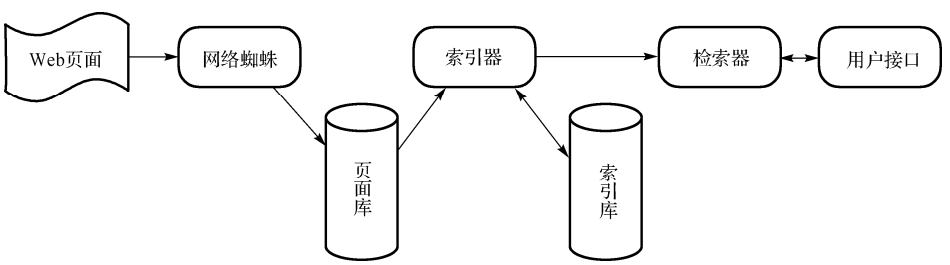


图 2-1　通用搜索引擎系统结构

（1）网络蜘蛛（亦称网络爬虫）在互联网中不断地搜索（也称爬行），发现和采集新的网页信息，然后将网页信息存入页面库，由索引器建立索引。

（2）索引器将分析网络蜘蛛所采集的信息，从中抽取出索引项，建立用于检索页面的索引表，存入索引库中。

（3）检索器将根据用户的查询请求和条件，从索引库中快速检索出网页信息，并通过网页相似度评价，对输出的结果进行排序。

（4）用户接口为用户提供一个输入查询请求和显示查询结果的用户界面。

3．通用搜索引擎不足

尽管通用搜索引擎已经成为搜索互联网信息的常用网络工具，但是它也存在一定的局限性：

（1）不同领域、不同背景的用户往往具有不同的检索目的和需求，通用搜索引擎所返回的结果往往包含用户并不关注的大量网页。

（2）通用搜索引擎的目标是尽可能高的网络覆盖率，有限的搜索引擎服务器资源与无限的网络数据资源之间存在一定的矛盾。

（3）通用搜索引擎主要提供基于关键字的信息检索，一般不支持基于语义的信息查询。

2.2.2　主题搜索引擎

主题搜索引擎是一种针对特定主题的搜索引擎，可以为某一特定领域、某一特定人群或某一特定需求提供信息检索服务，其特点就是“专、精、深”，与通用搜索引擎相比，主题搜索引擎显得更加专注、具体和深入。

1．主题搜索引擎特点

主题搜索引擎专注于特定主题或领域的信息搜索，对于非特定主题或领域的信息被视为无效信息。这就要求网络蜘蛛在互联网上搜集信息时，必须采用基于主题的搜索策略。网络蜘蛛按照预先设定的主题来搜集相关信息，减少了所采集的信息量，提高了索引库中的信息质量。

主题搜索引擎具有以下特点：

（1）领域范围小。由于专注于特定主题或领域，信息量相对较小，便于建立起一个专业信息收录全、能够实时更新的索引库，提高了信息的质量。

（2）词汇量小。只涉及某一个或几个主题或领域，能够降低词汇和用语的一词多义现象，而且利用主题词表进行规范和控制，提高了信息查全率。

（3）准确率高。可以通过专家指导等方式，提高查询语句的明确性和精确度，使查询结果的准确率大为提高。

（4）便于带宽的使用。信息采集量小，减少了网络传输量，有利于网络宽带的有效利用。

（5）查询响应时间短。由于索引数据库的规模小，有利于缩短查询响应时间，还可以采用复杂的查询语法，提高用户查询的准确度。

2．主题搜索引擎结构

主题搜索引擎是在通用搜索引擎结构上改进而成的，其改进主要表现为如下几个方面。

（1）主题网络蜘蛛。采用面向主题的网络蜘蛛，能够按照预先设定的主题有选择性地搜集相关的网页，排除不相关的网页。由于信息搜集范围要小得多，提高了信息搜集质量，并且便于信息实时更新。

（2）索引器。对搜集到的信息进行准确的分类标引是搜索引擎中最重要的一个环节。在通用搜索引擎中，对信息的分类标引主要有自动和人工两种。自动分类标引速度快，但精确度不高；人工分类标引精确度高，但速度太慢。而主题搜索引擎所涉及的领域和信息量都比较小，完全可以将两种分类标引方法有机地结合起来，在自动分类标引过程中加入人工智能，利用专家知识对信息进行分类标引，提高了信息质量。

（3）检索器及用户接口。对检索器和用户接口改进主要体现在 4 个方面：一是提供了充分表达用户查询要求的检索功能。二是提供了相似度反馈机制，不仅可以通过计算检索词在每个结果中的出现次数和出现位置来计算相似度，而且可以根据用户的要求决定相似度，通过多次交互逐步求精。三是能够集中处理检索，可通过站点共聚或内容共聚，减少信息总量，将检索结果集中处理成相同的格式反馈给用户。四是能够应用自动获得的领域模型和用户模型进行智能化信息搜集、索引和过滤，并自动地将用户感兴趣的有用信息提交给用户。

主题搜索引擎系统结构如图 2-2 所示，主要由面向主题的网络蜘蛛、索引数据库、关键词数据库、用户接口、关键词相似度计算、文档相似度计算、文档聚类器、检索器等部分组成。

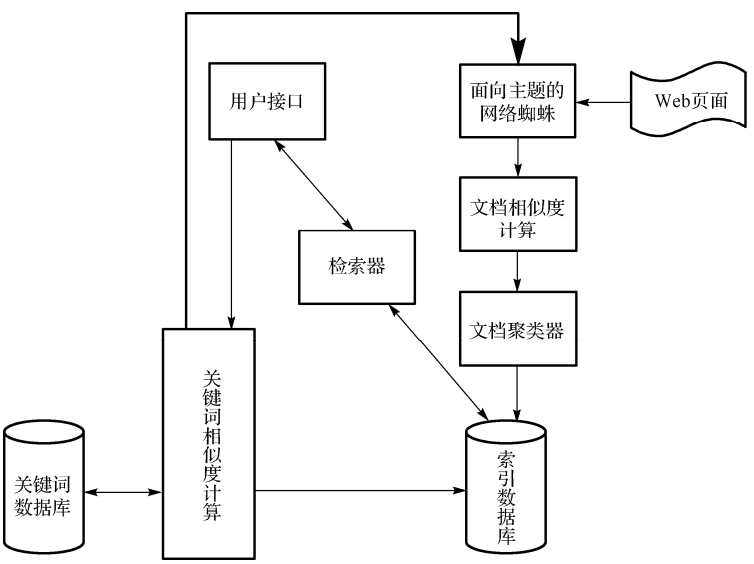


图 2-2　主题搜索引擎系统结构

主题搜索引擎工作过程如下：

（1）系统首先将人工收集到的常用关键词输入到关键词数据库中，然后启动面向主题的网络蜘蛛模块，根据关键词数据库中的关键词爬行 Web 页面，取回搜集到的文档。

（2）通过文档相似度计算模块计算其文档相似度，去除与主题无关的信息。然后通过文档聚类器模块将与主题相关的信息聚类成簇，并根据关键词建立索引，分类存入到索引数据库中。

（3）用户通过用户接口或用户界面输入相应的关键词，系统启动关键词相似度计算模块，查询关键词数据库中是否存在相匹配的信息，如果存在则直接从索引数据库中提取相关的信息建立索引。系统通过检索器模块，从索引数据库中快速检索出相关文档信息，并对文档信息与查询信息之间的相似度进行评价，以此来排序将要输出的结果，实现某种用户相关性反馈机制。

2.3　网络蜘蛛概念

2.3.1　基本概念

搜索引擎一直专注于提升用户的体验度，其用户体验度则反映在三个方面：准、全、快，即查准率、查全率和搜索速度，其中的难点在于准和全。搜索引擎的“准”，需要保证搜索到的前几十条结果都和搜索词密切相关；搜索引擎的“全”，则需保证不遗漏某些重要的信息，而且能找到最新的网页。这就需要搜索引擎有一个强大的网页搜集器，即网络蜘蛛。

网络蜘蛛是搜索引擎的核心部件，它的性能好坏直接影响到搜索引擎的整体性能和处理速度。在搜索引擎系统中，网络蜘蛛主要负责抓取网页、图片和文档等信息。其抓取过程是从给定的起始 URL 开始，沿着网页中的链接，按照一定的搜索策略进行遍历搜索，下载相应的网页，解析出网页中的超链接 URL，看是否已被访问过，将那些未访问过的 URL 加入到待爬行队列，然后再搜索其他链接指向的网页，循环往复。整个过程如同一个蜘蛛在蜘蛛网（Web）上爬行，因此网络蜘蛛也称网络爬虫。

通用搜索引擎所使用的网页搜集器为通用网络蜘蛛系统（简称通用蜘蛛），主题搜索引擎所使用的网页搜集器为面向主题的网络蜘蛛系统（简称主题蜘蛛）。

2.3.2　通用蜘蛛

1．通用蜘蛛组成结构

通用搜索引擎如同一个公共图书馆，通用蜘蛛就是这个图书馆的采购员，它试图满足各类用户的查询需求，所搜集的网页内容广而泛。通用蜘蛛系统结构如图 2-3 所示。

其中，各个主要模块的功能如下：

（1）爬行模块。该模块是网络蜘蛛和互联网的接口，主要作用是通过 HTTP 协议来完成对网页数据的采集，然后将采集到的网页提交给其他模块做进一步处理。

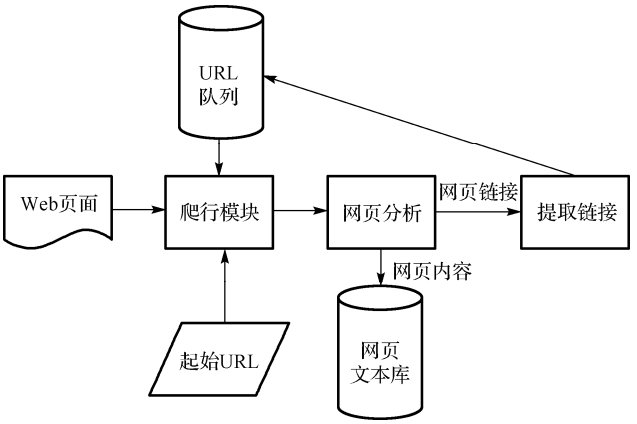


图 2-3　通用蜘蛛系统结构

（2）网页分析模块。该模块对所采集的网页进行分析，提取其中满足用户要求的 URL 链接，加入到 URL 队列中。如果网页链接中给出的 URL 是一个相对路径，则需要将相对路径转变为绝对路径。

（3）提取链接模块。该模块的主要作用是去除重复链接和循环链接。

（4）网页文本库。用来存储经过网页分析处理的网页，供后期处理使用。

（5）URL 队列。用来存放从网页中提取出的 URL，当 URL 队列为空时，网络蜘蛛程序停止爬行。

（6）起始 URL。提供种子 URL，用来启动网络蜘蛛程序。

2．通用蜘蛛工作原理

通用蜘蛛的目标就是尽可能多地采集网页，在这一过程中并不关注网页采集的顺序和被采集网页的相关主题。这就需要消耗较多的系统资源和网络带宽，并且这些资源的消耗并没有换来对所采集网页的较高利用率。

通用蜘蛛从网站某一个网页（通常是首页）开始，读取网页内容，并抽取出网页中的其他超链接地址，然后通过这些链接地址寻找下一个网页，这样一直循环下去，直到满足系统的停止条件，其工作流程如图 2-4 所示。

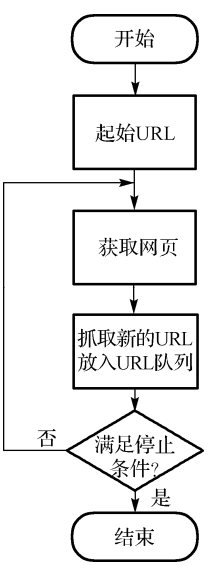


图 2-4　通用蜘蛛工作流程

3．通用蜘蛛搜索策略

通常，通用蜘蛛在抓取网页时采用两种搜索策略：深度优先搜索策略和广度优先搜索策略。

深度优先搜索策略是指网络蜘蛛从起始网页开始，一个链

接一个链接地跟踪下去，处理完这条路径之后再转入下一个起始网页，继续跟踪链接，直到遍历所有的网页及链接，搜索过程结束。这种搜索策略的优点是网络蜘蛛能够遍历一个 Web 站点或深层嵌套的文档集合。缺点是当 Web 站点的网页文件结构比较深时，有可能发生陷入进去而出不来的情况，即网络蜘蛛的陷入问题。

广度优先搜索策略是指网络蜘蛛从起始网页开始，首先搜索完一个网页中所有的链接，然后再继续搜索下一层，直到底层为止。这种搜索策略的优点是能够保证对浅层链接的优先处理，即使遇到一个深层分支时，也不会导致发生网络蜘蛛陷入深层文件中出不来的情况，并且能够在两个网页文件之间找到最短路径。广度优先搜索策略通常是网络蜘蛛的最佳策略，不仅容易实现，并且还能够实现并行处理，提高其抓取速度。然而，对于深层嵌套的网页文件集，广度优先搜索策略需要花费较长的时间才能搜索深层的网页文件。

两种搜索策略的区别可以用图 2-5 给出的示例来说明。

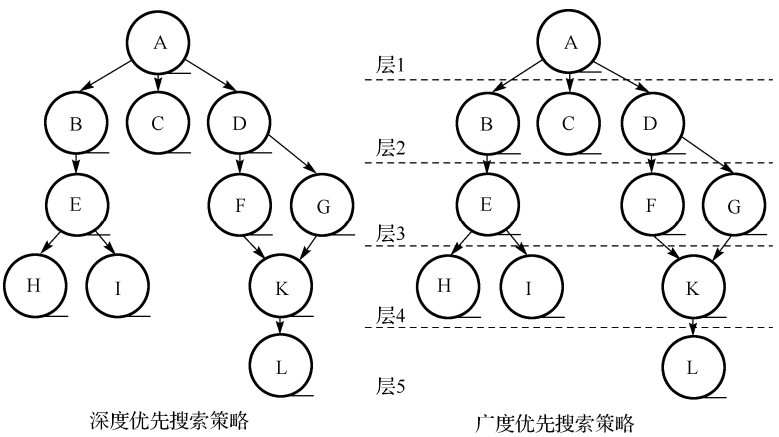


图 2-5　通用蜘蛛搜索策略

假设网络蜘蛛从 A 出发，两种搜索策略的搜索路径如表 2-1 所示，可见它们的搜索路径是不同的。

表 2-1　两种搜索策略的搜索路径

路径编号	深度优先搜索路径	广度优先搜索路径
1	A B E H	A
2	A B E I	B C D
3	A C	E F G
4	A D F K L	H I K
5	A D G K L	L

在主流的网络蜘蛛系统中，一般采用广度优先搜索策略为主、深度优先搜索策略为辅

的搜索策略。对于某些不被引用或很少被引用的网页，广度优先搜索策略可能会遗漏这些孤立的网页，而深度优先搜索策略可以搜索到这些网页。

2.3.3 主题蜘蛛

1．主题蜘蛛组成结构

主题搜索引擎如同一个专业图书馆，主题蜘蛛就是这个图书馆的采购员，它只需满足某一类用户的查询需求，只搜集与主题内容相关的网页。主题蜘蛛的组成结构如图 2-6 所示。

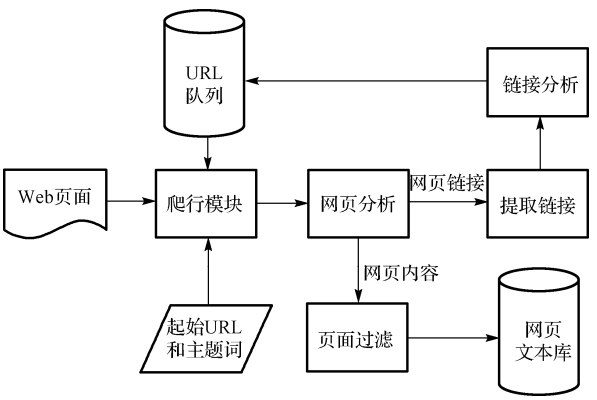


图 2-6　主题蜘蛛组成结构

主题蜘蛛主要由三个关键模块组成：

（1）网页分析模块。该模块采用了文本分类模型和算法对网页相关性进行分析。在爬行开始前，首先需要对样本网页信息进行学习，训练成一个主题相似度模型。当爬行到一个被认为与主题相关的网页后，将该网页提交给页面过滤模块进行主题相似度值计算，如果该网页相似度值大于或等于给定的阈值，则说明该网页与主题相关，则存入网页文本库；否则说明该网页与主题相关性小，丢弃该网页。

（2）链接分析模块。该模块是主题蜘蛛的核心模块，主要用于评估从主题相关网页中解析出来的 URL 链接与主题的相似度，并根据相似度值提供相应的搜索策略来指导网络蜘蛛的爬行过程。URL 链接的主题相似度值越大，爬行的优先级就越高，在 URL 队列中排列越靠前。反之，如果某一 URL 链接的主题相似度值小于给定的阈值，则将该 URL 链接及其所隐含的子链接一并去除。

（3）爬行模块。该模块是任何网络蜘蛛程序都不可缺少的通用模块。爬行模块首先从待爬行 URL 队列中取出排在首位的 URL，将该 URL 对应的网页抓取到本地，然后将该页面交给页面分析模块处理。在整个爬行过程中，爬行的次序和搜索策略都是由链接分析模块提供的。

2．主题蜘蛛工作原理

与通用蜘蛛不同的是，主题蜘蛛在爬行开始前需要就某个主题对样本网页信息进行学习和训练，建立相应主题相似度模型。在启动爬行后，首先从起始 URL 开始爬行，按照设定的搜索策略来搜索网页，对于所获取的网页，首先进行网页相关性分析，去除与主题不相关的网页；然后对所提取的 URL 进行链接相关性分析，设置该 URL 的优先级并存入 URL 队列。当满足停止条件时，爬行过程结束。其工作流程如图 2-7 所示。可见，存入网页文本库中的网页都是与主题相关的网页。

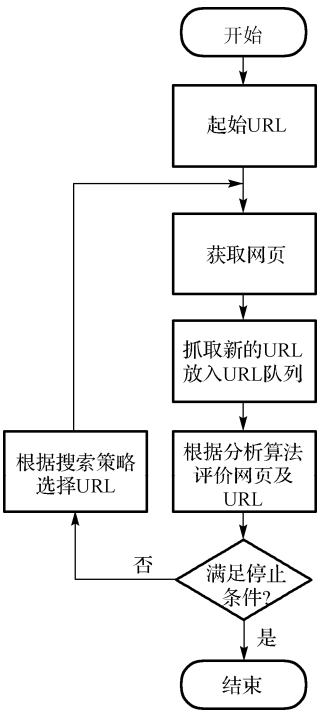


图 2-7　主题蜘蛛工作流程

3．主题蜘蛛搜索策略

由于主题信息一般只占整个 Web 空间很小的一部分，并且具有分散性，因此传统的深度优先搜索策略和广度优先搜索策略在 Web 信息搜集的效率上难以达到期望要求。由于主题蜘蛛的特点是采集的信息内容只限于特定的主题或专门的领域，因此在搜索信息过程中没有必要对整个 Web 空间进行遍历搜索，只需要选择与主题相关的网页进行访问即可。对主题搜索引擎而言，提高网络蜘蛛搜索效率的关键在于如何将不相关的网页快速地过滤掉，因为网页过滤的速度和准确性将会直接影响网络蜘蛛的性能。

主题蜘蛛在爬行过程中，如果对所发现的 URL 链接都不加选择的话，则大量的无关 URL 链接就会极大地浪费网络蜘蛛的处理时间。为了避免或减少这种现象的发生，主题蜘蛛必须对所发现的 URL 链接进行预测，只访问那些预测值符合给定阈值的 URL 链接。对于预测值符合要求的网页，系统还要对所抓取的网页进行主题相关性分析，以保证最终得到的网页是与主题相关的。

相比于通用蜘蛛，主题蜘蛛能够有效地发现主题相关的网页，并且能通过网页内容和链接结构来指导其资源发现过程。图 2-8 反映了通用蜘蛛和主题蜘蛛的爬行过程中的差异。

其中，浅色框代表主题无关的网页，黑色框代表主题相关的网页，虚线代表链接，箭头代表访问次序。

（1）通用蜘蛛以广度优先搜索策略，沿着每个链接进行爬行。假设从起始网页到目标网页需要爬行 i 步，那么在爬行到目标网页前必须先将 i–1 步内的网页爬行完。

（2）主题蜘蛛首先确定最有可能与主题相关的链接，忽略主题无关的网页。假设从起始网页到目标网页需要爬行 i 步，那么在爬行到目标网页前仅爬行 i–1 步内的网页的一个子集，在理想情况下，只要爬行 i 个链接就可以达到目标网页，这大大节省了爬行时间，提高了爬行效率。

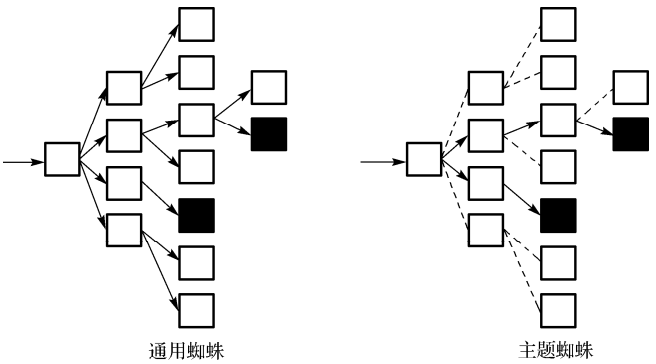


图 2-8　通用蜘蛛和主题蜘蛛的爬行区别

如果同一个主题相关的网页在 Web 网站上是平均分布的，则主题蜘蛛和通用蜘蛛在效率上没有明显的差别。经过大量的研究表明，Web 网站上的网页分布具有主题聚合性。利用这一特性，主题蜘蛛就可以优先对主题相关网页中的链接进行爬行。

与通用蜘蛛相比，主题蜘蛛更加专业化和可定制化。通用网络蜘蛛的目标是尽可能多地采集网页，在这一过程中它并不关注网页采集的顺序和被采集网页的主题，而主题蜘蛛能够定向性地采集与主题相关的网页，忽略无关的网页，并且还可以根据主题相似度值进行优先采集。因此，主题蜘蛛具有如下特点：

（1）抓取目标描述和定义。在主题蜘蛛采集网页之前，首先要确定一个抓取的主题。

（2）URL 搜索策略。它决定着待爬行 URL 的访问次序。通用蜘蛛以搜集海量数据为目标，在爬行时只需按照广度优先搜索策略或者深度优先搜索策略来抓取网页，尽量提高其爬行的范围。而主题蜘蛛因其自身的特点，在爬行过程中需要考虑优先访问与主题相关的网页，即考虑如何管理待爬行的 URL 队列，从而使得每次爬行都从相似度值最大的 URL 开始。

（3）网页相关性分析与过滤。对于已经下载的网页，需要通过适当的主题相似度模型和算法来判别该网页是否与主题相关。因此，主题相似度模型和算法的优劣将对网页和链接的主题相似度计算性能产生很大的影响。

2.4　网页搜索算法

2.4.1　网页特征选取

1．主题描述方式

主题蜘蛛需要对搜索目标的主题进行描述，主题描述方式可以分为基于目标网页特征、基于目标数据模式和基于领域概念等三种。

在基于目标网页特征的描述方式中，网络蜘蛛的搜索对象通常是网页，网页特征可以是网页的内容特征，也可以是网页的链接结构特征等。种子样本获取方式可以分为如下三种：

（1）预先给定的初始种子样本。

（2）预先给定的网页分类目录以及与分类目录对应的种子样本。

（3）通过用户行为确定的搜索目标样例，包括用户浏览过程中显示标注的样本、通过用户日志挖掘得到访问模式及相关样本。

在基于目标数据模式的描述方式中，网络蜘蛛的搜索对象是网页上的数据，其网页数据要符合一定的数据模式或者能够映射为目标数据模式。这类数据的典型代表就是电子商务网站的产品信息页面，具有统一的风格，其中的数据表示具有固定格式，并按照一定目录层次结构来组织。

在基于领域概念的描述方式中，通过建立目标领域的词典，从语义角度分析不同特征在某一主题中的重要程度。这种描述方式具有清晰的概念层次以及概念间、属性间的关系定义，能够方便地获得一个词语的同义词或上下义词。

2．特征选取方法

常用的特征选取方法有文档频率（Document Frequency，DF）、信息增益（Information Gain，IG）、互信息（Mutual Information，MI）、卡方检验（Chi-square /χ^2，CHI）等。

1）DF

DF 表示在训练集中包含某个特征项 t 的文档数。使用这种方法来衡量特征项重要程度是基于这样一个假设：DF 值较小的特征项对分类结果的影响较小。这种方法优先选取 DF 值较大的特征项，而 DF 值较小的特征项将被剔除，即特征项按照 DF 值排序。DF 是最简单的特征项选取方法，并且该方法的计算复杂度低，能够胜任大规模的分类任务。

2）IG

IG 通过统计某个特征项 t 在一个文档中出现或不出现的次数来预测文档的类别，IG 的计算公式如下：

$$G(t)=-\sum_{i=1}^{m}P_r(c_i)\lg P_r(c_i)+P_r(t)\sum_{i=1}^{m}P_r(c_i\mid t)\lg P_r(c_i\mid t)+P_r(\overline{t})\sum_{i=1}^{m}P_r(c_i\mid \overline{t})\lg P_r(c_i\mid \overline{t}) \qquad (2\text{-}1)$$

式中，$P_r(c_i)$表示一个文档属于类别 c_i 的概率；$P_r(t)$表示特征项 t 在一个文档内出现的概率；$P_r(\overline{t})$ 表示特征项 t 不在一个文档内出现的概率；$P_r(c_i|t)$表示特征项 t 在属于类别 c_i 的文档内出现的概率；$P_r(c_i\mid\overline{t})$ 表示特征项 t 不在属于类别 c_i 的文档内出现的概率。m 是文档类别数。$G(t)$值大则被选取的可能性大，即特征项按照 G 值排序。

3）MI

MI 使用如下公式计算某个特征项 t 和类别 c 之间的相关性。

$$I(t,c)=\lg\frac{A\times N}{(A+C)\times(A+B)} \tag{2-2}$$

式中，A 为 t 和 c 同时出现的次数；B 为 t 出现而 c 没有出现的次数；C 为 c 出现而 t 没有出现的次数。N 为所有文档数。如果 t 和 c 不相关，则 I （t, c）值为 0。如果有 m 个类别，每个 t 都会有 m 个值，取它们的平均值，就可以得到特征选取所需的一个线性序列。I 平均值大的特征被选取的可能性大。

4）CHI

使用 MI 衡量特征项的重要程度时，只考虑到了正相关对特征项重要程度的影响。如果特征项 t 和类别 c 反相关，则说明含有特征项 t 的文档不属于 c 的概率要大一些，这对于判断一个文档是否不属于类别 c 也是具有指导意义的。CHI 考虑了反相关性，使用如下公式计算特征项 t 和类别 c 的相关性。

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{2-3}$$

式中，A 为 t 和 c 同时出现的次数；B 为 t 出现而 c 没有出现的次数。C 为 c 出现而 t 没有出现的次数；D 为 t 和 c 同时没有出现的次数。N 为训练集中的文档数。与 MI 类似，如果 t 和 c 不相关，则 $\chi^2(t, c)$值为 0。与 MI 相同，如果有 m 个类别，每个 t 都会有 m 个值，取它们的平均，就可以得到特征选取所需的一个线性序列。χ^2 平均值大的特征被选取的可能性大。

2.4.2 网页搜索算法

为了高效地搜集与主题相关的网页资源，主题蜘蛛应尽可能多地搜集主题相关的网页，并尽可能少地搜集无关的网页，以确保搜集网页的质量。因此，研究人员提出了多种主题定制搜索策略和相关算法，下面介绍几种比较流行的搜索算法。

1．基于链接结构评价的搜索算法

基于链接结构评价的搜索策略是利用 Web 结构信息来指导搜索，并通过分析 Web 网页之间相互引用的关系来评价网页和链接的重要性。这种策略的基本思想来自于文献计量学的引文分析理论，将引文分析理论应用于Web环境时，主要采用基于链接结构的评价方法。下面介绍基于这种策略的常用两种搜索算法。

1）PageRank 算法

PageRank 算法是由 Google 公司研究人员提出并应用于搜索引擎中。该算法挖掘出网页间链接关系的价值，其挖掘过程也称为链接分析。简单地说，如果网页 A 链接到 B，则表示网页 A 的编写者对网页 B 的认可。或者说网页 A 为网页 B 投了一票，网页 B 的重要性被网页 A 认可。网页重要性可以从三个方面来评价。

（1）认可度越高的网页越重要，即反向链接越多的网页越重要。

（2）反向链接的源网页质量越高，被这些高质量网页链接指向的网页越重要。

（3）链接数越少的网页越重要。

PageRank 算法用于计算网页的重要性，对每个链入赋以不同的权值，链接提供的网页越重要，则该链入权值就越高，即当前网页的重要性是由其他网页的重要性决定的。PageRank 算法的计算公式如下：

$$PR_n(A)=(1-d)+d\times\left(\sum_{i=1}^{m}\frac{PR_{n-1}(T_i)}{C(T_i)}\right) \tag{2-4}$$

式中，$PR_n(A)$是网页 A 的级别（即 PageRank 值），$PR_{n-1}(T_i)$是网页 T_i 的级别，网页 T_i 存在指向网页 A 的链接，$C(T_i)$是网页 T_i 链出的链接数量，d 是阻尼系数，取值在 0～1，通过实验找出的最佳值为 0.85。

该算法不以站点排序，网页的级别由一个个独立的网页决定，即由链向网页的级别来决定，但每个链入网页贡献的值是不同的。如果 T_i 网页中链出越多，它对当前网页 A 的贡献就越小。网页 A 的链入网页越多，其网页的级别也越高。阻尼系数的使用，减少了其他网页对当前网页 A 的排序贡献。

该算法可以通过用户上网行为的随机冲浪模型来解释，随机冲浪模型对用户上网行为描述如下：

（1）用户随机选择一个网页作为上网的起始网页。

（2）看完这个网页后从该网页内所含的链接中随机选择一个网页继续浏览。

（3）沿着链接继续浏览，直到对某个主题感到厌倦而重新随机选择另一个网页浏览。如此反复，直到结束。

随机冲浪模型将用户点击链接的行为视为一种不关心内容的随机行为，而用户单击网页内链接的概率完全由网页上链接数量的多少来决定，这也是式（2-4）中 $PR_{n-1}(T_i)/C(T_i)$的原因。一个网页通过随机冲浪模型到达的概率就是链入它的网页上链接被点击概率之和。阻尼系数 d 的引入是因为用户不可能无限地点击链接，常常因劳累而随机跳入另一个网页。d 可以视为用户无限地点击下去的概率，$1-d$ 则就是网页本身所具有的网页级别。

PageRank 算法主要考虑了链接的结构特征，而忽略了网页内容与主题的相关性，容易出现采集偏离主题的“主题漂移”问题。因此，研究人员提出了一些改进的 PageRank 算法，如在 PageRank 算法中考虑了用户从一个网页直接跳转到非直接相邻的但是内容相关的另一个网页的情况等。

2）HITS 算法

PageRank 算法对于链出的权值贡献是平均的，也就是不考虑不同链接的重要性。而 Web 链接却具有以下特征：

（1）有些链接只起导航或广告作用，还有些链接具有注释性，只有注释性的链接才用于权威判断。

（2）基于商业或竞争因素考虑，很少有 Web 网页指向其竞争领域的权威（Authority）网页。

（3）权威网页很少具有显式的描述，比如 Google 主页不会明确给出 Web 搜索引擎之类的描述信息。因此，权值的平均分布不符合链接的实际情况。

HITS（Hyperlink-Induced Topic Search）算法定义了另一种网页，称为中心（Hub）网页。Hub 网页是提供指向 Authority 网页链接集合的网页，它本身可能并不重要，或者说没有几个网页指向它，但是 Hub 网页的确提供了指向某个主题的重要站点链接集合，例如一个课程主页上的推荐参考文献列表。一般来说，好的 Hub 网页指向很多好的 Authority 网页，好的 Authority 网页是有很多好的 Hub 网页指向的网页。这种 Hub 与 Authority 网页之间的相互加强关系，可用于 Authority 网页以及 Web 结构与资源的自动发现，这就是 Hub/Authority 方法的基本思想。

HITS 算法是基于 Hub/Authority 方法的搜索算法，算法如下：

（1）将查询 q 提交给基于关键字匹配的搜索引擎。搜索引擎返回很多的网页，从中取前 n 个网页作为根集，用 S 表示。S 满足以下条件：S 中网页数量相对较小；S 中大多数网页是与查询 q 相关的网页；S 中网页包含较多的 Authority 网页。

（2）通过向 S 中加入被 S 引用的网页和引用 S 的网页，将 S 扩展成一个更大的集合 T。

（3）以 T 中的 Hub 网页为顶点集 V_1，以 Authority 网页为顶点集 V_2，V_1 中的网页到 V_2 中的网页的链接为边集 E，形成一个二分有向图 $SG=(V_1, V_2, E)$。对 V_1 中的任何一个顶点 v，用 $h(v)$表示网页 v 的 Hub 值，对 V_2 中的顶点 u，用 $a(u)$表示网页的 Authority 值。开始时 $h(v)=a(u)=1$，对 u 执行 I 操作修改它的 $a(u)$，对 v 执行 O 操作修改它的 $h(v)$，然后规范化 $a(u)$和 $h(v)$。如此重复计算 I、O 操作，直到 $a(u)$和 $h(v)$收敛。

I 操作为：

$$a(u)=\sum_{v:(v,u)\in E} h(v) \tag{2-5}$$

O 操作为：

$$h(v)=\sum_{v:(v,u)\in E} a(u) \tag{2-6}$$

每次迭代后需要对 $a(u)$和 $h(v)$进行规范化处理：

$$a(u)=\frac{a(u)}{\sqrt{\sum_{q\in V_2} a(q)^2}}$$

$$h(v)=\frac{h(v)}{\sqrt{\sum_{q\in V_2}a(q)^2}} \tag{2-7}$$

I 操作反映了如果一个网页有很多好的 Hub 指向，则权威值会相应地增加，即权威值增加为所有指向它的网页的现有 Hub 值之和。*O* 操作反映了如果一个网页指向很多好的 Authority 网页，则 Hub 值也会相应地增加，即 Hub 值增加为该网页链接的所有网页的权威值之和。

PageRank 和 HITS 两种算法的共同点是利用网页之间的引用关系来确定链接的重要性，其优点是考虑了链接的结构特征，但也存在一些缺陷：一是忽略了网页与主题的相关性，在某些情况下，会出现搜索偏离主题的“主题漂移”问题，二是在搜索过程中需要重复计算 PageRank 值或 Authority 与 Hub 权值，计算复杂度随访问网页和链接数量的增长呈指数级增长。

2．基于网页内容评价的搜索算法

基于网页内容评价的搜索策略是利用网页文本内容作为领域知识指导搜索，并根据网页文本与主题之间相似度的大小来评价链接价值的高低。下面介绍基于这种策略的两种常用搜索算法。

1）Fish Search 算法

Fish Search 算法也称为鱼群搜索算法，它将在网络上遍历搜索的网络蜘蛛比喻为海里的一个鱼群，当鱼群找到大量的食物（主题相关网页）之后，这些鱼就变得强壮，并繁殖更多的后代；反之，鱼群就变得虚弱，后代也少。当鱼群找不到食物（无相关网页）或者水被污染（带宽不够）的时候，鱼群就会死亡。该算法的关键是根据用户的种子站点和查询的关键词或短语，将包含查询字符串的网页看作与主题相关，计算该网页与主题的相似度，动态地维护待爬行 URL 的优先级队列 url_queue。

当获取一个网页后，提取该网页所有的 URL，这些 URL 所对应的网页称为孩子网页。如果获取的网页与主题相关，将孩子网页的深度设置成一个预先定义的值，否则将孩子网页的深度设置成一个小于父亲网页深度的值。当这个深度为零时，这个方向的搜索就停止。

按照下列启发策略，将深度大于 0 的孩子网页的 URL 加入到 url_queue 的顶部：

（1）相关网页的前 $\alpha\times$width 个孩子加入到 url_queue 的顶部，其中 α 是预设的大于 1 的常量。

（2）无关的网页的前 width 个孩子 URL 加入到 url_queue 队列中紧靠着相关网页的孩子节点后面。

（3）剩下的孩子 URL 加入到 url_queue 的尾部，它们只有在时间允许的情况下才有可能被搜索。

上述的三种情况可以用一个变量 potential_score 来等价描述。在情况（1）下，变量设

置为 1；在情况（2）下，变量设置为 0.5；在情况（3）下，变量设置为 0。待爬行 URL 队列按照该变量值来排序，算法伪代码如算法 2-1 所示。

算法 2-1　Fish Search 算法伪代码

```
Fish_Search(Starting_URLs,topic,width,D)
{
     enqueue(url_queue,Starting_URLs,D);              //将种子 URL 入栈，深度为 D
     int numVisited = 0;
     while(numVisited<MAX_PAGES && number != 0){
          (url,depth)= dequeue_top_link(url_queue);
          page = crawl_page(url);
          numVisited++;
          enqueue(crawled_queue,url);
          url_list = Extract_link(page);
          sim_score = sim(topic,page);                //计算相似度，判断当前节点是否相关
          enqueue(buffered_page,sim_score);           //将结果保留
          if (depth>0){
STEP1:    if (当前网页不相关){
                对 url_list 的前 width 个孩子节点(Child_node), potential_score=0.5;
                对剩余孩子节点，potential_score = 0;
                }
                    else{
                      对url_list 的前(α×width) 个孩子节点(α 为预先设置的常量，一般为1.5)potential_score = 1;
                      对所有剩余的孩子节点，potential_score = 0;
                      }
STEP2:  for(each u in url_list){
                    if(u In url_queue){
                      比较 url_queue 中的 score 和 u 的 score，用最大值取代 url_queue 中的 score;
                      如果有需要，按照 score 对 url_list 排序;
                    }else
                         如果有需要，按照对 url_list 寻找合适位置插入;
                    }
STEP3:     for(each u in url_list){                      //计算深度 depth
                    if(当前网页相关)
                         depth(u)= D;
                    else
                         depth(u)= depth(page)- 1;
                    if(u in url_queue)
                         比较 url_queue 中的深度和 depth(u),用最大值取代 url_queue 中的 depth;
                  }
               }//end of if (depth>0)
          }//end of while
}
```

该算法是一种基于客户端的搜索算法，其优点是模式简单、动态搜索。但也存在一些缺点，例如只使用简单的字符串匹配来分配 potential_score 的值，并且只有 1、0.5、0 三个值，分配的值不能完全代表与主题的相似度；在 url_queue 中，优先级值之间的差别太小，当很多的 URL 具有相同的优先级，并且在搜索时间受到限制时，可能将后面更重要的网页忽略了；使用 width 参数来调节删除网页后面的 URL 个数也不尽合理，有可能导致丢掉很重要的资源。

2）Shark Search 算法

Shark Search 算法是对 Fish Search 算法的一种改进，主要改进了网页与查询信息相似度计算方法和 potential_score 值计算方法，具体改进如下：

（1）在网页与查询信息的相似度计算中引入了向量空间模型，对相似度值进行细化，使其取值在 0～1，而 Fish Search 算法的相似度计算为简单的两值判断，不够细致和精确。

（2）在网页与查询信息的相似度计算中考虑了链接附近的文字（如锚文本及其上下文）所包含的提示信息，使相似度计算更加准确。

由于在孩子节点的 potential_score 计算中综合考虑了上述两个因素，因此提高了相似度计算的准确性。

基于网页内容评价的搜索策略是根据语义相似度的大小来决定链接的访问顺序，其优点是计算量比较小。然而，由于 Web 网页不同于传统的文本，它是一种半结构化的文档，其中包含了许多结构信息，Web 网页不是单独存在的，网页中的链接在一定程度上反映了网页之间存在着某些关系。采用这种搜索策略的网络蜘蛛忽略了这些信息，因此在链接预测和利用方面存在一些缺陷，容易造成网页的误选。

2.4.3　链接分级搜索

由于网络论坛、博客等互动式网站是网络舆论的主要来源，网民通过在这类网站上发布帖子来表达自己的意见和观点，容易形成网络舆情。因此这类网站成为网络舆情监测和信息采集的主要对象。由于这类网站有别于一般的静态网络，将传统的搜索策略直接用于网络论坛、博客网站信息采集时往往达不到令人满意的效果，需要根据这类网站的特点，采取有针对性的搜索策略。

1．网络论坛、博客网站特点

通过分析各种网络论坛、博客网站自身的结构特点以及链接结构，可以总结出网络论坛、博客网站具有如下特点：

（1）链接种类众多。除了一般用户所关心的文章或帖子对应的链接之外，论坛、博客中还存在着大量的功能性链接和噪声链接。所谓功能性链接是指为了完成某一功能或操作而

设置的链接，例如“登录”、“评论”等；而噪声链接是指广告链接之类与用户所关心的文章完全无关的链接。

（2）文章中链接所处层次不固定。有些文章链接可能置于某一目录式板块的索引页，甚至网络论坛、博客网站的首页之上；而另一些文章则置于某一特定的分档之中，如博客网站中博主自己的文章档案链接之下，有效链接所处层次不固定这一现象在博客网站中尤其突出。

（3）链接冗余现象显著。链接冗余是指同一网页与多个链接相对应的现象。这一现象普遍存在于网络论坛、博客网站之中，例如一篇文章中可能引用了一个链接，但是另外一篇文章中可能也引用同一链接，但链接的 URL 在形式上很可能是完全不同的。

如果网络蜘蛛没有针对网络论坛、博客网站自身的结构特点采用相应的搜索策略，则有可能导致如下方面的问题：一是大量的无效链接被采集；二是由于有效链接往往位于不同的深度层次，影响到采集覆盖率；三是大量的链接冗余现象有可能导致网络蜘蛛陷入采集陷阱之中。

2．链接分级模型

对于网络论坛、博客网站中各种形式的链接，可以将它们抽象成如下的链接类型。

（1）文章链接：网络论坛、博客网站中有效帖子等形式的文章所对应的链接，每一篇文章都有其唯一的链接，该类链接称为文章链接。

（2）博主链接：博客网站一般包含许多注册博主，每一博主有其唯一的链接，该类链接称为博主链接。

（3）板块链接：论坛网站一般可以划分出若干板块，如新闻板块、人文板块、体育板块等，每一板块有其唯一的链接，该类链接称为板块链接。

（4）目录链接：由于博主链接与板块链接比较相似，并且它们都有一个显著特征，即它们所对应的网页中都包含了若干文章链接，因此可以把两者抽象为目录链接，即目录链接 = 博主链接∪板块链接。

（5）其他链接：将一些功能性链接、噪声链接等与采集主题无关的一类链接称为其他链接。

上述链接类型基本覆盖了网络论坛、博客网站中所有形式的链接。一般情况下，目录链接中往往包含若干文章链接，这些文章链接则是网络蜘蛛所需要采集的重点链接，而其他链接则是网络蜘蛛所应规避的链接。

在此基础之上，可以把网络论坛、博客网站中的链接归纳划分为三种级别：目录链接、文章链接、其他链接。这三种级别的链接基本覆盖了网络论坛、博客网站中所有的链接形式，如图 2-9 所示。

从图 2-9 可以看出，网络蜘蛛所关心的文章链接往往包含在一个目录链接之下。因此，可以将上述构成抽象成为网络论坛、博客网站的链接分级模型，即网络论坛、博客网站中包含的大量文章链接都抽象归纳于一个目录链接之下，如图 2-10 所示。

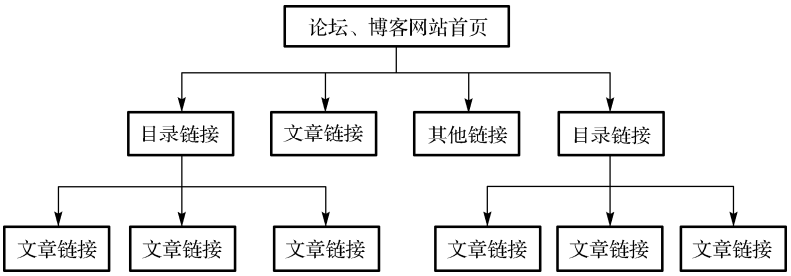


图 2-9　网络论坛、博客网站的链接构成

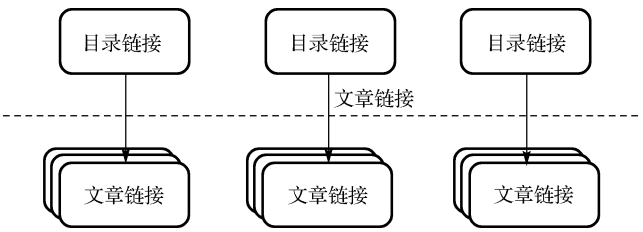


图 2-10　网络论坛、博客网站的链接分级模型

3．“目录链接-文章链接”搜索策略

根据链接分级模型，网络蜘蛛首先搜索目录链接，去除其他链接的干扰后，再采集目录链接下的文章链接，有助于提高网络蜘蛛的搜索效率。这种搜索方式称为“目录链接-文章链接”搜索策略。

“目录链接-文章链接”搜索策略具有如下优点：

（1）该策略可以使网络蜘蛛的注意力集中在目录链接与文章链接上，从而能够规避无效链接的干扰。

（2）该策略不同于广度优先策略或深度优先策略，在逻辑上规定了两级深度，并限制了各个文章链接之间的干扰，在搜索过程中能够避免搜索陷阱现象的发生。

（3）该策略更加符合人的思维方式，可以融入人工智能方法，提高网络蜘蛛智能化水平，进一步提高搜索效率。

因此，网络蜘蛛在采集网络论坛、博客网站信息时，比较适合采用链接分级模型进行搜索，能够有效地提高搜索效率。

2.5　网页相似度计算

在主题蜘蛛中，需要对采集下来的每一个网页进行内容分析，判断它是否与所要采集

的主题内容相关。在网页相似度计算中，采用向量空间模型来表示网页文本内容，该模型具有算法简单、计算复杂度低等特点，比较适合对网页文本内容进行实时处理。

2.5.1 向量空间模型

为了实现对文本的计算机处理，需要采用适当的表示模型来描述文本内容，通过从文本中提取某些特征词对文本进行量化，为相似度计算等后续的文本处理提供基础。

常用的文本表示模型有向量空间模型（Vector Space Model，VSM）和语言模型（Language Model，LM），其中，向量空间模型使用一组从文档中提取的特征项（T_1，T_2，T_3，…，T_n）来表示文档，特征项是指用来表示文档内容特征的基本语言单位，如字、词、词组或短语等，通常是文档中最能够反映文本基本特征的词，也称为特征词。每个特征项根据文本处理任务的不同，提取的特征项也可能有所不同。对于每一个特征项 T_i，根据它在文本中的重要程度赋予一个权值 W_i。这样，一个文本的特征项集合就可以看作是一个 n 维的坐标系，W_1，W_2，W_3，…，W_n 为对应的坐标值。这样，通过向量空间模型表示之后，一个文档集合就变成了一个矩阵，每一行代表一个文档，每一列代表这个文档中的某个特征项。这样构成的矩阵具有很高的维数，需要经过降维处理后才能做进一步的处理。

在信息查询处理中，提取用户查询请求和文档中的特征项构成向量空间，根据向量空间的相似度大小来排列查询结果。使用向量空间模型，按照特征词的维度分别对查询词和文档进行向量化，然后采用适当的相似度度量方法（如余弦系数法等）来计算文档与查询词之间的相似度，从而优先检索那些与查询词相似度大的文档，并且能够按照与查询词的相似度对检索出的文档进行排序。向量空间模型不仅可以方便地产生有效的查询效果，而且还能提供相关文档的文摘，对查询结果进行分类，为用户提供准确定位所需的信息。向量空间模型如图 2-11 所示。

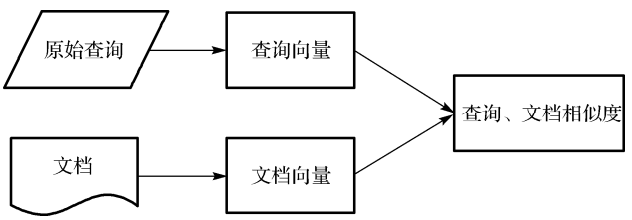


图 2-11　向量空间模型

向量空间模型的基本思想如下：

（1）把原始查询和文档都看作文本，使用同样的向量化过程分别得到查询向量和文档向量。

（2）采用适当的相似度度量方法来计算查询向量和文档向量之间的相似度。

（3）按照与查询词的相似度值大小对检索结果进行排序。

（4）根据检索结果，做进一步的相关性反馈。

下面详细介绍如何把文档内容简化为特征项及其权值的向量表示，以及如何把对文档内容的处理简化为向量空间的向量计算。

2.5.2　相似度计算

1．文档空间向量表示

文档可以有多种类型，比如 Web 网页、段落和句子等，索引项则可能是汉字、词、词组和短语等。

首先对一个文档进行分词处理，并去除那些停用词。然后对剩余的词进行合并处理，生成该文档的各个索引项。在一个给定的集合中，首先对每个文档进行以上操作，获得每个文档索引项的集合。然后再把所有文档的索引项合并，形成了一个代表整个文档集合的索引项集合，整个索引项集合表示了一个“空间”，在这个“空间”中的每一个索引项都有一个权值。假如每个文档都可以用一个索引项的集合来表示，那么这个集合称为“文档空间”。在一个文档空间中，可以给每一个索引项赋一个权值，表示这个索引项在该文档中的重要性。一个索引项在不同的文档空间中的权值是不同的，表示这个索引项在不同文档中的作用是不同的。如果一个索引项在一个文档空间中的值为 0，则表示这个索引项和该文档没有关系。如表 2-2 所示，文档 1 的向量空间是（1,4,5），文档 2 的向量空间是（4,9,2），文档 3 的向量空间是（0,0,10）。

与文档空间相对应的一个概念是“项空间”，它是指一个索引项在文档集合中的各个文档中权值的集合。如表 2-2 所示，索引项 T_1 的向量空间是（1,4,0），索引项 T_2 的向量空间是（4,9,0），索引项 T_3 的向量空间是（5,2,10）。

表 2-2　索引项在文档中的权值表

	索引项 T_1	索引项 T_2	索引项 T_3
文档 1（D_1）	1	4	5
文档 2（D_2）	4	9	2
文档 3（D_3）	0	0	10

如果把索引项空间和文档空间合并在一起，就构成了一个文档-索引项的矩阵。若有 n 个索引项，文档 D_i 就可以表示为一个 n 维向量，$W_{i,j}$ 表示文档 D_i 的第 j 维的权值，矩阵如图 2-12 所示。

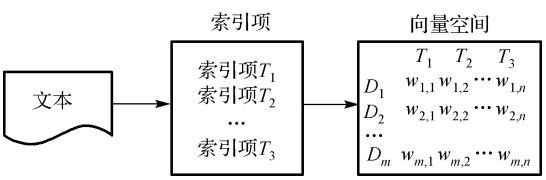


图 2-12　文档和索引项的向量表示

2．向量间相似度计算

相似度度量方法有很多，最常用的是余弦系数法，余弦系数法采用计算向量之间夹角余弦的方法来计算检索向量和文档向量的相似度。设检索向量 q =（q_1, q_2, …, q_n），文档向

量 $d=(d_1, d_2, \cdots, d_n)$，余弦系数计算公式如下：

$$\cos(q,d)=\frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n} q_i^2 \sum_{i=1}^{n} d_i^2}} \tag{2-8}$$

两个向量间的夹角越小，夹角余弦值越大，其相似度越高。例如，对表 2-2 中文档 $d_1=(1,4,5)$，文档 $d_2=(4,9,2)$，文档 $d_3=(0,0,10)$；检索条件 $q=(1,0,2)$，则可以用以上公式计算每个文档和检索条件的夹角余弦值，计算结果如下：

$$\cos(q,d_1)=\cos(\theta_1)=\frac{1\times1+4\times0+5\times2}{\sqrt{(1^2+2^2)\times(1^2+4^2+5^2)}}=\frac{11}{\sqrt{5\times42}}=0.759$$

$$\cos(q,d_2)=\cos(\theta_2)=\frac{4\times1+9\times0+2\times2}{\sqrt{(1^2+2^2)\times(4^2+9^2+2^2)}}=\frac{8}{\sqrt{5\times101}}=0.356$$

$$\cos(q,d_3)=\cos(\theta_3)=\frac{0\times1+0\times0+10\times2}{\sqrt{(1^2+2^2)\times10^2}}=\frac{20}{\sqrt{5\times100}}=0.894$$

以上结果表明，d_3 的相似度 $>d_1$ 的相似度 $>d_2$ 的相似度。

3．索引项的权值

1）索引项的选择

对于汉语而言，可以选择字、词、短语，甚至句子或句群作为索引项。索引项的选择要由处理速度、精度、存储空间等方面的具体要求来决定。由于词汇是文本中最基本的表示项，在文本中出现的频率较高，并呈现一定的统计规律，因此将词或词组作为索引项比较合适。需要注意的是，要将文本中一些没有实际意义，但使用频率很高的虚词和功能词组成停用词表，把文档中的停用词先滤除后再处理。

2）索引项权值计算

在每个文档中，每个索引项都有一个权值，表示该索引项在一个文档中的重要程度，即一个索引项在多大程度上可以将这个文档与其他文档区别开来。

由于文档的数量庞大，因此需要通过算法来计算每个索引项的权值。比较常用的权值计算方法是项频度－逆向文档频度加权法（TF-IDF）。

项频度 *tf* 是指项 t_i 在文档 d_j 中的出现次数，记作 $tf_{i,j}$。$tf_{i,j}$ 值越高，意味着项 t_i 对于文档 d_j 就越重要。文档频度 *df* 是指含有项 t_i 的文档数量，记作 df_i。df_i 值越高，意味着项 t_i 在衡量文档之间相似度方面的作用越低。逆向文档频度 *idf* 则是表示了一个项在整个文档集合中的特性，用来衡量这个项在整个文档的分布情况。一般而言，*idf* 和 *df* 为反比关系，如下式所示：

$$idf_i=\lg\left(\frac{N}{df_i}\right) \tag{2-9}$$

式中，N 为文档集合中的文档数目。idf_i 值越高，意味着项 t_i 对于文档的区别意义越大。如果一个索引项仅出现在一个文档中，$idf = \lg N$；如果一个索引项出现在所有的文档中，$idf = \lg 1 = 0$。

3）索引项权值的加权

索引项权值的加权过程就是给那些经常出现在一个文档中，而不常出现在其他文档中的项赋予更高的权值，即让“特别的词”从“一般的词”中凸现出来。

加权的方法有两种，一种是对项频度 tf 进行加权，方法有最大法、扩展法和对数法等；另一种是对项向量的长度（文档长度）进行加权处理，主要有余弦系数法和支点法等。最常用的索引项权值计算方法是余弦系数法，即则索引项 i 的权值公式为：

$$w_i = \frac{\sum_{i=1}^{n} tf_i \times \lg\left(\frac{N}{df_i}\right)}{\sqrt{\sum_{i=1}^{n} (tf_i)^2 \times \lg^2\left(\frac{N}{df_i}\right)}} \tag{2-10}$$

4．中文分词

分词就是把一个句子按照其中词的含义进行切分，将连续的字串或序列按照一定的规范重新组合成词序列。与英文不同，汉语中最小的单位不是词，而是字，但具有一定语义的最小单位却是词。在英文的行文中，单词之间是以空格作为自然分界符的，在词理解上就比较直观。而汉语只有在句与句之间才通过标点或段落来简单划界，词与词之间没有这样的分界符。例如，英文句子“I am a student”，汉语意思是“我是一个学生”。在英文文本处理中，计算机可以通过空格和标点来确定“student”是一个词。但在中文文本处理中，让计算机理解“学”和“生”是一个词就不太容易。汉语分词就是要把汉语的这种序列切分成有意义的词，以便机器理解。

常用的分词方法一共有三种：正向最大匹配分词、逆向最大匹配分词和基于统计的词网格分词。

（1）正向最大匹配分词。正向最大匹配方法的基本思想是，假设自动分词词典中的最长词条所含汉字个数为 I，则取被处理材料当前字符串序数中的 I 个字作为匹配字段，查找分词词典。若词典中有这样的一个 I 字词，则匹配成功，匹配字段作为一个词被切分出来；如果词典中找不到这样的一个 I 字词，则匹配失败。匹配字段去掉最后一个汉字，剩下的字符作为新的匹配字段，再进行新的匹配，如此下去，直至切分成功为止。

（2）逆向最大匹配分词。逆向最大匹配分词方法的分词过程与正向最大匹配分词方法相同，不过它是从句子（或文章）的末尾开始处理，每次匹配不成功时去掉的是最前面的一个汉字。

（3）基于统计的词网格分词。词网格分词的第一步是选择词网格构造，利用词典匹配，列举输入句子所有可能的切分词语，并以词网格形式保存。实际上，词网格是一个有向

无环图，它包含了输入句子所有可能的切分，其中的每一条路径代表一种切分。第二步，选择计算词网格中的每一条路径的权值，权值通过计算图中每一个节点（词）的一元统计概率和节点之间的二元统计概率的相关信息得到，然后根据图搜索算法在图中找到一条权值最优的路径，对应的路径即为最后的分词结果。由于词网格分词方法是基于统计的方法，因此具有比较高的分词正确率和较好的可扩充性，并且可以通过加入相应的统计信息来扩展不同的功能。

2.6 主题蜘蛛组成

网络信息采集是网络舆情分析的基础，网络信息搜集的质量直接关系到网络舆情分析的效果。因此，网络信息搜集工具功能和性能是非常关键的。下面介绍一种主题蜘蛛的系统结构和组成原理。

2.6.1 系统结构

主题蜘蛛在搜索过程中需要完成三个基本功能，一是网页读取，主要功能是读取 Web 服务器上的网页内容，并对其进行分词处理，以便形成特征向量进行主题相似度计算；二是链接提取，主要功能是分析网页中的链接，将网页上的所有链接提取出来，并计算每个链接的相似度，将与主题相关的链接存入到待爬行 URL 队列中；三是内容提取，主要功能是分析网页内容，将网页中所有标记语言符号去掉，只留下网页文字内容。根据功能需求，一个主题蜘蛛系统通常由 4 个主要功能模块组成，其系统结构如图 2-13 所示。

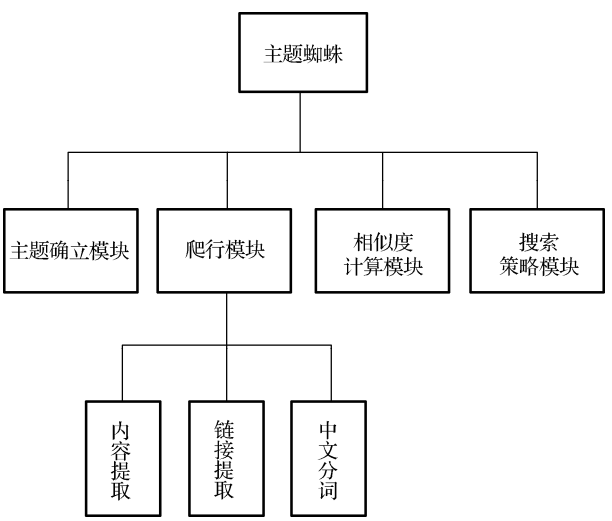


图 2-13　主题蜘蛛系统结构

2.6.2　主题确立模块

主题确立模块的作用是描述和确立所要搜索的主题，这是主题蜘蛛工作的基础。可以用一组特征词来确立主题，其中每个特征词可以定义不同的权值。权值定义有两种方法：手工设置和特征提取。特征提取是指给定一个与主题相关的网页集合，由程序自动提取这些网页的共同特征，并根据其特征出现的频率来确定权值，特征提取的优点是权值量化精确，但要求所选取的网页集合必须具有代表性和概括性，否则就可能出现较大的偏差。手工设置的优点是实现简单，具有一定的准确性，与实际情况不会出现大的偏差，缺点是权值的量化不够精确。

在系统实现中，可以采用手工设置为主，特征提取为辅的方式来确立主题。首先手工设置一组关键词并分配权值，然后将这组关键词应用到搜索引擎中查找出对应的网页，从中选择若干相关的网页（称为样本网页）作为训练集合，去除掉网页的 HTML 标记得到网页正文内容，然后将它们串联成一篇文档，经中文分词处理后进行特征提取，文档中的一个段落就表示一个样本网页的正文内容。

特征提取是对给定样本网页进行分析，自动提取这些网页中的关键词，并根据关键词在网页中出现的次数计算其权值，最后综合各个样本网页的关键词，确定一组能够代表主题的特征词。选出的特征词应当满足如下条件：（1）完全性，能体现样本网页所包含的内容；（2）区分性，通过特征词能确定搜索网页是否与样本网页相关。特征词的权值采用 TF-IDF 方法来计算，TF-IDF 方法的思想是：一个特征词在同一文档中出现的次数越多，说明它区分文档内容属性的能力越强，其权值应越大；而在若干文档中均出现的特征词，说明它区分文档内容属性的能力越低，其权值应越小。在实际使用中，可以采用式（2-11）的变形公式来计算特征词在文档中的权值，即：

$$w_{ik}=\frac{t_{ik}\lg\left(\dfrac{N}{n_k}+0.1\right)}{\sqrt{\sum_{k=1}^{n}(t_{ik})^2\times\lg^2\left(\dfrac{N}{n_k}+0.1\right)}} \tag{2-11}$$

式中，t_{ik} 表示特征词 t_k 在文档 d_i 中的出现次数，N 表示文档数据库中全部文档的总数，n_k 表示文档数据库中含有特征词 t_k 的文档数。

将特征提取选出的特征词与手工设置的关键词进行合并，根据实际情况调整最终的特征词个数及权值。由此确立的主题就是一个能够代表主题相关文档的基准文档向量，向量的维数为特征词的个数，每一维分量的大小为每个特征词的权值。

2.6.3　爬行模块

爬行模块可进一步分为内容提取、链接提取和中文分词等三个子模块。

1．内容提取子模块

内容提取子模块的作用是对所抓取的 HTML 网页进行语法分析，提取出标题、网页正文、链接及其他相关内容，为后续的主题相似度计算提供基础。

HTML 网页有 5 种定义好的组件：文本、注释、简单标签、开始标签和结束标签。

（1）文本就是在 HTML 网页上看到的词句。除了脚本代码外，HTML 文档中的所有数据，只要不是标签的组成部分，都被认为是文本。文本是格式化的，并且受包围它的标签控制。例如，<h1>Hello World! <h1>，这段 HTML 代码包含两个标签和文本“Hello World！”。如果数据位于文本之外，则不会看作文本。在这种情况下，它被当作脚本代码。例如 JavaScript 程序包含在标签<script>和</script>之间，这样就确保了蜘蛛程序不会将 JavaScript 代码与文本数据相混淆。

（2）注释表示 HTML 网页中不会显示给用户的那部分内容。它们通常是 HTML 程序员所留下的注释性说明。由于注释不会显示，对用户是不可见的，因此本系统在解析 HTML 页面时，自动忽略注释。

（3）简单标签是指完全单个表示的 HTML 标签，它们没有相应的结束标签。简单标签主要用来控制显示格式和美化界面。

（4）开始标签和结束标签非常像简单标签，两者之间的唯一区别是：开始标签有一个相应的结束标签，结束标签出现在后面。开始标签和结束标签用来控制其所包含的 HTML 代码的功能。

HTML 网页是半结构化的，可以从中提取出代表该网页的 4 种基本属性：锚文本、标题、正文、超链接。

（1）锚文本：除了网页标题可以描述网页之外，还可以用一些锚文本来描述它。例如，门户网站主页可能被另外一些网页中存在的锚所指向，其锚文本便是该网页的最佳描述。特别是某些没有标题的网页，锚文本是有益的补充。

（2）标题：这里的标题特指 HTML 标识语言中<title></title>中间的文字部分，这部分文字表达了网页的基本含义。与锚文本相同的是，都是用来描述网页的内容的属性，而与锚文本不同的是，这个标题是由该网页制作者来编写的。

（3）正文：锚文本、标题都是网页的简短描述，而正文是一个网页的主体内容，它完整地描述了网页的主体内容。

（4）超链接：超链接（简称链接）是从一个网页指向一个目标的连接关系，这个目标可以是另一个网页，也可以是相同网页上的不同位置，还可以是一个图片、一个电子邮件地址、一个文件，甚至是一个应用程序。而超链接的对象，可以是一段文本或者是一个图片。

在系统中，对每一个采集下来的网页都要进行提取处理。锚文本通常出现在标记<a…>…的尾部，提取出的锚文本主要用于系统中主题相关性链接的预测；网页的正文部分一般出现在<body>和</body>标记之间，只要找出这两个标记就可以找到正文所在的位

置，然后去除这两个标记之间的其他标记内容就得到正文文本，将文本内容进行分词处理后，就可以参与向量空间模型的计算，对网页内容的主题相似度进行判别，若相关则存储，否则将被删除；网页中的超链接先要由系统的搜索策略进行相关性预测，符合要求的将被加入到待爬行 URL 队列，不符合要求的将被丢弃。经过处理并被存储的网页是以文本文档形式保存的。

2．链接提取子模块

链接提取子模块的作用是提取出网页中所包含的链接。蜘蛛程序在爬行过程中必须能够从一个网页移动到另一个网页，这是通过分析每个网页上的 HTML 代码，查找网页中所有链接到其他网页的标签来实现的。大多数链接到其他网页的标签是使用一种称为超文本链接（HREF）的特殊属性来标识的。网页使用 HREF 链接在一起，HREF 是指定其他网页链接的 HTML 属性。

所有的 HTML 链接均包含在 HTML 的 HREF 属性中，HREF 不是一个 HTML 标签，它只是一个属性，总是和其他的锚标签结合在一起使用，锚标签的功能是指向蜘蛛程序要访问的另一个网页。对于蜘蛛程序来说，下面的锚标签只表示有另外一个名为 nextpage.html 的网页需要解析，其他数据都将被忽略。

Click Here

HREF 包含的 URL 路径可以是绝对的，也可以是相对的。绝对 URL 指定了一个准确、无歧义的网络资源位置，包含了主机名和文件名，例如：http://www.websiteA.com/intro.html，其中 www.websiteA.com 是主机名，intro.html 是文件名。相对 URL 仅指定绝对 URL 的一部分。根据 HREF 所包含的数据，URL 链接有三种类型：内部链接、外部链接和其他链接。

（1）内部链接：它指向的网页与包含该链接的网页在同一网站或 Web 服务器上。图 2-14 给出了一个示例网站的内部结构。该网站由很多内部相连的网页构成，其主页包含了指向其他 4 个网页的内部链接，其版块列表网页又包含指向 6 个网页的内部链接。

（2）外部链接：它指向的网页所在的 Web 网站与包含该链接的 Web 网站不同。例如，网站 *A* 中某个网页的链接指向了另一个网站 *B*，则这个链接就称为外部链接。因为当用户点击这个链接时，浏览器会让他们进入网站 *B* 的网页，这是网站 *A* 之外的网页。

（3）其他链接：该链接是不指向网页的链接。它仅在指向 E-mail 地址或其他资源时才是有效的。指向不属于 HTTP 模式的链接属于此类。例如，mailto 模式可以用来指定 E-mail 地址。

对于上述三种链接，主题蜘蛛的抓取对象是内部链接，因为外部链接将导致蜘蛛程序不断地访问新 Web 网站而进入无法终止状态。因此，蜘蛛程序在抓取网页中所包含的链接时，需要判别其 URL 中主机名是否与该网页的网站名相匹配，例如，在抓取新浪网页中所包含的链接时，其 URL 中必须包含有“sina”字符串。

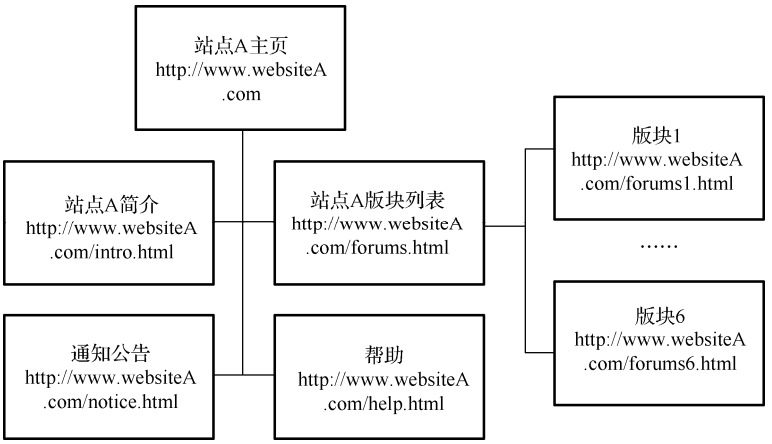


图 2-14　一个示例网站结构

蜘蛛程序在抓取每个 URL 链接时，首先判断该链接所指向的网页是否为 html 或是 htm 或是 shtml 文件，否则将丢弃该网页。对于符合要求的文件，则按顺序读取文件，遇到超链接标记如<a href = …>等，则提取其中的 URL 链接，并从成对的该标记之间提取锚文本作为该链接的说明文字。重复上述过程直到文件结束。然后将提取出来的 URL 链接按照绝对 URL 的格式补充完整，并检查该 URL 链接是否已经存在于待爬行 URL 队列中，只将新的链接存入待爬行 URL 队列，而丢弃已有的链接，从而保持待爬行 URL 队列中链接的唯一性。

3．中文分词子模块

中文分词子模块的作用是对提取的每个网页正文进行分词处理，切分出能够代表该文档内容的特征词或关键词。然后将切分出的特征词作为文档向量，与预先确定的主题向量进行相似度计算，以判断该网页是否为主题相关网页。

在中文分词处理中采用一种中文处理软件来实现，该软件是一个支持二次开发的开源软件包，为用户提供了许多接口和方法，包括分词接口和关键词分析接口等，具有分词准确率高、效率高、接口方便灵活等特点，并且能够在程序的开始和结束时对分词组件进行初始化和卸载，满足了实时处理网页中文分词的应用要求。

4．爬行效率

蜘蛛程序的任务是高效、稳定地抓取网页数据，其爬行效率是衡量一个蜘蛛程序可用性的重要指标。为了提高其爬行效率，蜘蛛程序通常采用多线程技术并发地抓取网页数据，并且用户可以根据网络带宽大小、机器性能高低等实际情况来设置和调整线程的数量。

使用多线程抓取网页时，需要考虑线程间同步问题。例如，系统使用多线程对待爬行 URL 队列进行操作，某个线程在访问该队列时，要保证其他线程不能同时对该队列进行提取、添加、判重、判空、排序等操作。临界区是一种常用的线程同步技术，可以将待爬行

URL 队列当作临界区资源，每个线程从待爬行 URL 队列中提取链接时，首先锁定该队列，然后选择队列头的 URL 来爬行，选择好 URL 后将队列解锁，以便其他线程访问该队列。当对该队列进行添加、判重、判空、排序等操作时，也要将队列锁定，待操作完成后再将队列解锁。

2.6.4　相似度计算模块

相似度计算模块的作用是计算网页内容与主题之间的相似度。在网页主题相似度计算时，首先使用向量空间模型来表示文本，然后使用某种相似度度量方法来计算其相似度，这里使用基于余弦系数法的相似度度量方法，即计算两个空间向量之间的夹角余弦，算法的实现步骤如下：

（1）将确立的主题看作由一组特征词构成，把特征词的个数 n 作为向量空间的维数，每个特征词的权值 W_i 作为每一维分量的大小，则主题用向量表示为：

$$W=[W_1,W_2,\cdots,W_n]$$

（2）对每一个采集到的网页进行内容分析，经过中文分词等处理后形成一个代表网页内容的关键词集合，并按照 TF-IDF 公式计算关键词权值，将其映射到 n 维向量空间中。因此对于每一个采集到的网页，可以用一个相应的空间向量来表示：

$$D_i=[D_{i1},\cdots,D_{ij},\cdots,D_{in}]$$

式中，D_i 表示第 i 个网页的空间向量，D_{ij} 表示第 i 个空间向量中第 j 维分量的大小，即第 i 个网页中第 j 个关键词的权值。

（3）计算网页与主题的相似度，即空间向量$[D_i]$与$[W]$之间的相似度，可以简化为计算向量空间中两个向量之间的夹角余弦或者向量之间的距离来度量。夹角越小，夹角余弦越大，距离越短，就表明两个向量之间的相似度越高。余弦系数法计算公式参见式（2-8），这里的空间向量分别是 W 与 D_i。将计算结果与预设的网页相似度阈值 r_1 比较，当相似度 $(W, D_i)\geqslant r_1$ 时，则可以认为该网页与主题是相关的，将该网页保存到网页库中；否则是不相关的，将该网页删除。r_1 的取值需要根据经验和实际要求来确定，r_1 值设置较小，则获取的网页较多；r_1 值设置较大，则获取的网页就较少。

2.6.5　搜索策略模块

搜索策略模块的作用是计算每个链接的相似度，并按每个链接相似度大小来指导其爬行过程，使整个网页采集过程能够持续下载相关网页。这里的链接相似度是指 URL 所指向的网页与指定主题的相关程度，相似度越高就越有可能是相关网页，在爬行时将按照相似度值大小来选择待爬行 URL 队列中的链接。

在计算链接相似度时，首先利用已访问网页的主题相似度和网页链接的上下文信息来

计算每个链接的相似度，然后将计算结果与预设的链接相似度阈值相比较，只有大于相似度阈值的链接才能被加入到待爬行 URL 队列中，并按相似度值大小来排序 URL，系统将优先处理相似度值大的 URL。

2.6.6 系统界面

一个主题蜘蛛系统运行的初始界面如图 2-15 所示。在系统采集过程中，该界面可以显示网页链接的下载情况、被下载网页数目和被拒绝下载网页数目等信息。

图 2-15　系统运行的初始界面

点击“开始抓取”按钮进入到抓取参数设置界面，用户可以根据实际需要来设置起始 URL、网页保存目录、预设网页相关度和预设 URL 相关度等参数，如图 2-16 所示。

图 2-16　抓取参数设置界面

设置好系统参数后，可以启动系统运行，运行时的界面如图 2-17 所示。系统在运行过程中，用户可以根据需要单击对应按钮执行暂停或停止操作。

图 2-17　系统运行界面

第3章

微博网络信息传播机制

3.1 引言

微博是一种集成化、开放式的互联网社交服务平台，用户通过 140 字以内的微博发布信息，实现即时分享。此外，用户可以根据自己的兴趣偏好，选择关注其他用户，构建自己的关注网络。

2006 年 3 月，博客技术的创始人威廉姆斯所创建的互联网公司Obvious开发并推出了 Twitter 网站。Twitter的出现把人们带入了一个全新的互联网时代，即微博时代。关于名字 Twitter 的来历，其英文原意为鸟儿的叽叽喳喳声，创始人认为鸟儿的叫声具有短、频、快的特点，符合该网站的内涵，因此选择了 Twitter 作为网站的名称。在最初阶段，Twitter 只提供向好友的手机发送文本信息的服务，后来逐渐提供一些新的服务，比如，用户可以通过 SMS、电子邮件、Twitter网站或 Twitter 客户端软件（如 Twitterrific）接收和发送信息，现在的 Twitter网站已发展成一个集社交网络和微博为一体的综合社交服务网站。

此后，国内外出现了大量类似 Twitter 的网站，国外的有Plurk、Jaiku等。国内的有饭否、做啥、叽歪、嘀咕、贫嘴、同学网、腾讯滔滔、9911 等，其中，饭否影响力较大。2009 年上半年，饭否的用户从年初的 30 万激增到 100 万，随着众多文化名人的加入以及国内众多知名媒体开辟饭否官方账号，饭否一度成为中国微博的标杆。后来，国内的四大门户网站均开设了微博网站，微博用户数量迅猛地增长。尽管近几年受到微信等即时通信工具的冲击，但微博的网民数仍然是比较庞大的。

Twitter 网站是世界上率先推出的微博平台，以崭新的信息交流方式在世界上引起极大的反响，成为全球影响力最大的微博平台。国内著名的微博平台有新浪微博、腾讯微博、搜狐微博等，其中新浪微博是国内最大的微博平台，其注册用户数量超过 5 亿人，日活跃用户数达到 4 620 多万人。

微博打破了传统媒体单一的舆论主导权，给大众提供了一个自由发表意见并与他人分享的平台，在一定程度上保证了公众的话语权。因此，微博极大地解放了公众话语权，促使了公众话语权的回归，开创了一个平民化的信息传播模式。

微博作为一种特殊的社交网络，用户不但可以有选择地连接感兴趣的用户，关注其信息，而且也可以被其他用户相互连接，交流信息，具有社交网络和媒体网络的双重特性，一些学者认为微博是一个社交媒体网络。

微博作为新兴的社交媒体，越来越受到重视。在国外，许多政治人物、政府部门、新闻机构等都开通了 Twitter 账号，作为与民众沟通交流、获取信息的手段。在国内，自 2009 年云南省政府新闻办开设了国内第一家政府微博“微博云南”后，全国各地的政府部门都陆续开通了政务微博，实时发布消息，与民众互动。国内的 CCTV、人民日报、新华通讯社等主流媒体也都在新浪等微博平台上开通了官方微博，作为新闻发布、了解民意、监测舆情的主要渠道。根据人民网微博舆情监测室对 2015 年突发公共事件舆情的统计，在各种突发公共事件的舆论关注度中，微博通常达到万条以上，而网络新闻、论坛帖子通常在几百条到几千条。可见，微博已经成为网络舆情的主要来源地。

微博的广泛应用也引起了国内外学术界的关注，研究人员对微博网络的基本特性、网络结构、信息传播、用户行为等方面进行了广泛研究。通过对微博用户转发行为特性的分析，寻找其中的内在规律，不仅可以为网络舆情监测、突发事件预测等提供科学依据，还可以为商家分析用户的购买喜好、推荐商品以及精准投放广告等提供有益的帮助。

本章主要介绍微博网络的用户转发特性、转发行为预测、转发峰值分析、意见领袖识别等内容。

3.2　微博用户转发特性

微博转发是微博网络提供的一种信息传播机制，用户可以将关注者发布的微博转发到自身平台上，然后分享给粉丝。通过这种信息传播机制，使得微博能够在更大范围内传播和分享。可见，用户转发行为是推动微博信息传播的重要因素。

用户转发行为与用户关系类型有关。在微博网络中，用户之间存在三种社会纽带关系，即强连接、弱连接及权威连接，不同的社会纽带关系对用户转发行为的影响也不同。因此，在分析用户转发行为时，首先需要识别用户之间的社会纽带关系。

在识别社会纽带关系时，首先需要提取出权威比率、微网络结构、地理距离以及性别等特征，然后采用适当的模型来分析各个特征间的相关性，并根据相关性来分析用户转发行为的内在动力。

3.2.1　转发行为特性

在微博网络中，对于信息的传播，每个用户扮演两种不同的角色，一个是接收者，转发来自于其他用户发布的微博；另一个是发布者，用户发布自己的微博。

对于微博用户转发行为，主要从接收者角度来分析，也就是说主要考虑用户是否会转

发其他用户的微博，而不考虑其微博是否被其他用户所转发。例如，对于用户 A 来说，仅考虑用户 A 是否会转发其关注者 B 的微博，而不考虑用户 A 发布的微博是否会被其他用户或粉丝转发。从本质上来说，从接收者或发布者的角度来研究微博转发行为是一样的，只是用户角度不同而已，从接收者角度来研究只是为了简化问题。

在微博网络中，定义 A→B 为一条关注边，用户 A 关注用户 B（或者用户 B 被用户 A 所关注）。在该关注边中，用户 A 是用户 B 的粉丝或用户 B 是用户 A 的关注者。成为用户的粉丝意味着能够自动地接收该用户发布的微博。因此，在关注边 A→B 中，用户 A 能够接收到用户 B 的所有微博。同时用户 A 可以转发用户 B 的微博，通过这一功能，用户 A 的粉丝能够自动地接收到该微博，并可以再次转发。因此，转发功能实现了微博信息在网络中的传播。但是网络关注边方向与微博传播的方向是相反的，例如在 A→B 中，表示用户 A 关注用户 B，而微博则是由用户 B 转发到用户 A 的。

图 3-1 给出微博用户关注网络图，Bob 关注了 Greg、Harry、Fred 以及 Carrol 等，同时被 Alice、Dave、Greg 以及 Harry 等关注，如图 3-1 中的不同箭头方向，其中存在一部分用户与 Bob 相互关注，例如 Greg 与 Harry。对于 Bob 而言，将会转发被他关注的用户的微博。

对于关注边 A→B，什么因素导致用户 A 转发用户 B 的微博。由于用户之间存在着不同类型的社会纽带关系，不同类型的社会纽带关系将会导致不同的微博转发行为，因此需要从社会纽带关系来研究微博转发行为。

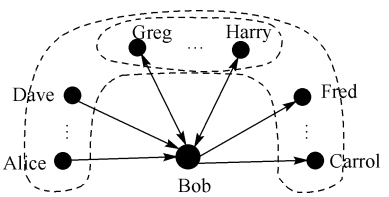


图 3-1　微博网络用户关系网络示例图

在微博网络中，用户之间存在三类社会纽带关系：强连接、弱连接和权威连接。其中，强连接关系是指用户之间具有高度的互动，在某些存在的互动关系形态上较为亲密；弱连接相对于强连接关系而言，用户之间并不具有高度的互动，但能够传递非重复性的信息；权威连接完全不同于强连接和弱连接，主要表现在用户之间的非对称性，非对称性包括两个方面，一是用户影响力的非对称，例如名人的影响力比一般用户大很多；二是信息传播的非对称性，例如名人的微博很容易被一般用户转发，而一般用户的微博很难被权威用户转发。在权威连接关系中，信息传播方向一般由权威高的用户到权威低的用户。在社会科学中，这种现象称为服从权威。

显然，用户之间不同的社会纽带关系将导致不同的微博转发行为。但是在微博网络中，用户之间的社会纽带关系是很难发现的，也很难定义用户之间是否存在强连接、弱连接或权威连接等关系。因此，如何识别出不同的社会纽带关系是一个关键性问题。

在社交网络中信息传播存在两个重要进程，即同化与社会影响。同化是指信息在网络传播过程中容易导致用户和与自己的观点、价值观相似的用户建立连接关系，最终使社交网络的结构发生变化。社会影响是指信息在网络传播过程中导致相邻用户的观点、价值观等属性逐渐趋于一致，最终使两个用户具有相似性。因此，在信息传播过程中，不同社会纽带关

系的用户将受到同化和社会影响的影响，最终导致了网络结构和用户属性的变化，也就是说，不同的社会纽带关系将具有不同的网络结构和用户属性。反过来说，通过提取网络结构和用户属性等特征，就能够识别出用户之间的社会纽带关系。网络结构特征可以从微博用户关注图得到，包括权威比率、微网络结构；而用户属性则可以从用户的个人资料中提取出来。下面将介绍相关特征的提取方法。

1．权威连接比率

在社交网络中，关注边 A→B 上的两个用户在不同的社会纽带关系中有着不同的社会权威。在强连接或弱连接中，用户大多是朋友、同事等对等关系，社会地位是平等的。而在权威连接中，名人与普通用户的社会地位并不平等，名人往往比普通用户有更高的社会地位和权威。因此，关注边中不同的社会权威比率可以间接地反映出两个用户是否存在着强连接、弱连接或权威连接关系。由于一个用户的粉丝数反映了用户的影响力，因此可以采用用户的粉丝数来刻画用户权威。对于关注边 A→B，两个用户权威比率定义如下。

$$P_{\mathrm{A}\to\mathrm{B}}=\frac{\#\mathrm{Follower(B)}}{\#\mathrm{Follower(A)}} \tag{3-1}$$

式（3-1）反映了关注边 A→B 中两个用户的权威差异性，$P_{\mathrm{A}\to\mathrm{B}}$ 值越大表明两个用户的社会地位越不平等，则关注边 A→B 为权威连接的可能性越大。

$P_{\mathrm{A}\to\mathrm{B}}$ 是一个连续值，需要对其离散化，信息熵是最常用的离散化度量方法。基于熵的离散化是一种监督的、自顶向下的分裂方法，在计算和确定分裂点（划分属性区间的数据值）时利用类分布信息，选择具有最少熵的属性作为分裂点，使区间划分离散化。

将 $P_{\mathrm{A}\to\mathrm{B}}$ 离散化为三大类：Small、Medium 和 Large。对 $P_{\mathrm{A}\to\mathrm{B}}$ 信息熵离散化后的结果如表 3-1 所示。

表 3-1　权威比率离散化分布

属　　性	值
Small	$P_{\mathrm{A}\to\mathrm{B}}<4$
Medium	$4<P_{\mathrm{A}\to\mathrm{B}}<100$
Large	$100<P_{\mathrm{A}\to\mathrm{B}}$

2．微网络结构

在微博网络中，用户关注关系将构成一个有向网络图，而具有不同社会纽带关系的用户有着不同的网络结构，这里主要关注用户间微网络结构。相比于全局网络结构，微网络结构更能够反映出用户之间的社会纽带关系。例如，在权威连接中，用户与其关注者往往是单向关系，通常是普通用户单向地关注权威人士而权威人士并不会关注普通用户。然而在强连接或者弱连接中，用户与其关注者更容易相互关注，形成双向关系。为了区分双向关注边 A↔B 是否为强连接或弱连接，引入第三方邻居用户，即是否存在与用户 A、B 都相互关注

的用户。如果用户 A、B 间的社会纽带关系越强烈，则与他们都相互关注的用户数就越多。例如，在朋友圈子中，两个用户关系越紧密，则他们共同拥有的朋友数量就越多。

下面通过三种不同类型的微网络结构来区分不同类型的社会纽带关系，如图 3-2 所示。在 Pattern I 中，用户 A、B 相互关注，同时至少存在一个第三方用户与用户 A、B 也两两相互关注，即存在强连接的三角关系；在 Pattern II 中，用户 A、B 相互关注，与 Pattern I 不同之处在于，不存在一个第三方用户与用户 A、B 两两相互关注；在 Pattern III 中, 用户 A、B 是单向关注关系。这三种不同类型的微网络结构反映了不同的社会纽带关系，强连接的用户容易形成 Pattern I 微网络结构，弱连接的用户容易形成 Pattern II 微网络结构，而权威连接的用户则容易形成 Pattern III 微网络结构。

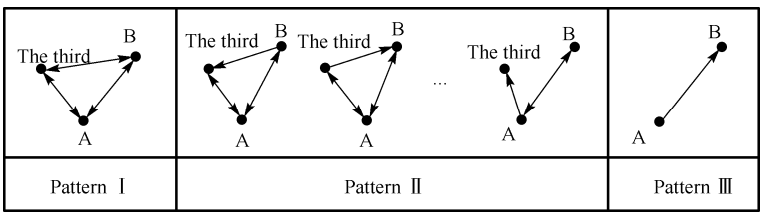


图 3-2　三种不同类型的微网络结构

3．地理距离

用户之间居住的地理距离对用户的社会纽带关系也有着不同的影响，因此地理距离也可以用来作为区分社会纽带关系的特征。例如，在强连接关系中，经常见面相处、距离较近的两个用户，容易建立起这种关系。如果两个用户距离较远，维持这种关系的成本将会大大增加，而不容易建立起这种关系。因此，建立强连接关系常常受到地理距离的限制，距离较近的两个用户比距离较远的两个用户更加容易建立强连接关系。而在权威连接关系中，普通用户可以通过一些公共媒体平台单向地关注权威人士，维持这种关系较为容易，由此可见，地理距离对权威连接的建立影响较小。

利用用户资料中所填写的省份与城市来定义两个用户的地理距离的远近，细分为三类：Near、Medium 和 Far，它们分别被定义为关注边 A→B 中两个用户处在相同城市、处在不同城市但处在相同的省份、处在不同的省份。

4．用户性别

用户之间不同的性别也有可能预示着不同的社会纽带关系。例如，相对于异性，同性关系的用户之间更加容易建立朋友关系。另外，社会科学家研究也表明，不同性别用户的思想观念、价值取向以及思维方式等并不相同，因此对于相同的微博，他们的兴趣也不一样，需要考虑不同性别对转发行为的影响。对于关注边的两个用户，可以分为四类：MM、MF、FM 和 FF，其中 MM 表示男性用户 A 关注男性用户 B；MF 表示男性用户 A 关注女性

用户 B；FM 表示女性用户 A 关注男性用户 B；FF 表示女性用户 A 关注女性用户 B。

以上所有特征都将影响着用户的微博转发行为，表 3-2 给出所有特征及相应的值，其中圆括号中的索引值代表相应特征值，例如，在权威比率特征中，索引值 1、2、3 分别代表相应的 Small、Medium 和 Large 值，其他类似。

表 3-2　微博转发特征以及相应值

	属　　性	值
A	权威比率	Small，Medium，Large （1，2，3）
B	微网络结构	Pattern I，Pattern II，Pattern III（1，2，3）
C	地理距离	Near，Medium，Far （1，2，3）
D	性别	MM，MF，FM，FF （1，2，3，4）
E	转发	Yes，No（1，0）

3.2.2　转发特性分析

下面通过实验数据对微博用户转发行为特性进行分析。

1．实验数据集

实验数据集来源于新浪微博。数据集采集于 2011 年 5 月至 7 月，随机选取了 3 430 个种子用户以及其关注的 171 769 个用户，然后排除不活跃的关注用户，数据集最终收集到 702 789 条活跃关注边，其中 185 327 条边含有转发记录。

不活跃的关注用户是指在关注边中存在不活跃的一些关注用户，他们几乎从不发布微博，很难有转发行为，因此需要从数据集中排除不活跃的关注用户。对于关注用户的活跃程度，采用下式来定义：

$$\theta = \frac{T}{R} \tag{3-2}$$

式中，T 和 R 分别表示发布微博的数目和关注用户注册时间的长短，θ 表示关注用户活跃程度，θ 值越大，则此用户越活跃，反之亦然。对于关注者的活跃阈值，设定 $\theta=2$，即当 θ 大于 2，视为活跃用户，否则为不活跃用户。

2．微博数据分布

表 3-3 给出了所有活跃关注边的分布，其中特征及相应的值如表 3-2 所示。在表 3-3 中，特征 A 对应于权威比率，相应索引值 1、2、3 分别代表 Small、Medium 和 Large，每个单元格的值对应于关注边 A→B 相应属性值的个数。例如，表 3-3（a）的最左上角单元格的值为 1059，则表示权威比率为 Small、微网络结构为 Pattern I、地理距离为 Near、性别为 MM 以及有转发记录的活跃关注边的数量是 1 059 条。

表 3-3 （a）微博转发数据分布Ⅰ

		D	1						2					
		C	1		2		3		1		2		3	
A	B	E	1	0	1	0	1	0	1	0	1	0	1	0
1	1		1 059	3 363	1 663	5 804	3 816	22 441	771	3 594	1 391	5 991	4 463	19 044
	2		35	333	49	636	364	3 841	44	610	107	1132	447	5 062
	3		254	2 027	441	4 305	1 704	19 840	186	3 096	306	5 122	1 343	22 679
2	1		621	982	1 036	1 735	2 178	6 765	145	473	316	996	760	3 460
	2		32	75	51	189	323	1 036	287	51	45	140	168	727
	3		1 230	47 093	2 736	10 917	12 400	47 840	435	2 270	1 392	5 574	5 708	24 438
3	1		24	37	79	65	229	214	17	9	19	22	45	93
	2		1	4	4	9	31	34	6	3	1	3	11	24
	3		1 525	2 239	4 722	6 421	23 209	31 536	988	1 226	3 142	3 623	19 132	21 533

表 3-3 （b）微博转发数据分布Ⅱ

		D	3						4					
		C	1		2		3		1		2		3	
A	B	E	1	0	1	0	1	0	1	0	1	0	1	0
1	1		590	1 919	912	3 404	2 409	11 772	1 029	2 548	1 455	4 284	4 551	16 147
	2		29	298	41	636	316	2 844	28	320	67	718	304	3 945
	3		80	809	187	1 693	732	10 374	112	1 192	236	2 644	1 811	16 417
2	1		241	525	469	1 037	1 260	4 168	233	319	341	693	816	2 661
	2		6	70	40	191	188	909	3	68	25	116	151	640
	3		463	2 109	1 563	5 666	7 800	27 443	326	1 411	980	3 501	5 854	19 016
3	1		156	24	52	48	68	138	19	8	62	20	79	62
	2		0	2	3	7	33	50	0	3	0	5	24	24
	3		1 142	1 454	2 772	4 269	20 647	25 913	1 135	864	2 570	2 929	19 685	21 428

图 3-3 给出了所有活跃关注边 A→B 中用户关注边数的分布，它符合幂律分布，在大部分关注边中两用户权威差值较少，而小部分关注边的两用户权威差值较大，这说明大部分关注边都是地位相称的用户。

图 3-4 给出了所有关注边 A→B 中用户转发次数的分布，它也符合幂律分布，大部分用户的微博几乎没有或者很少被转发，而小部分用户的微博则被大量转发，尤其存在少数用户转发数高达 100 次以上。由此可见，虽然用户关注很多其他用户，但是真正感兴趣的只是其中极少部分用户。

3．转发行为特性分析

下面讨论权威比率与微网络结构、地理距离与微网络结构、性别与微网络结构等两两组合属性对转发因子的权重，属性组合的权重越大，表示对转发的影响越大。

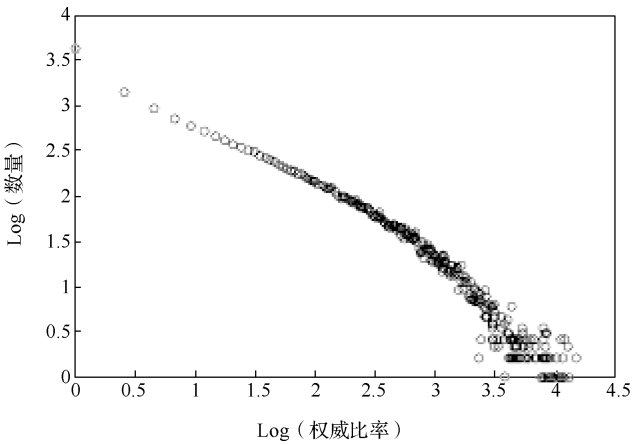


图 3-3　用户的关注边数分布

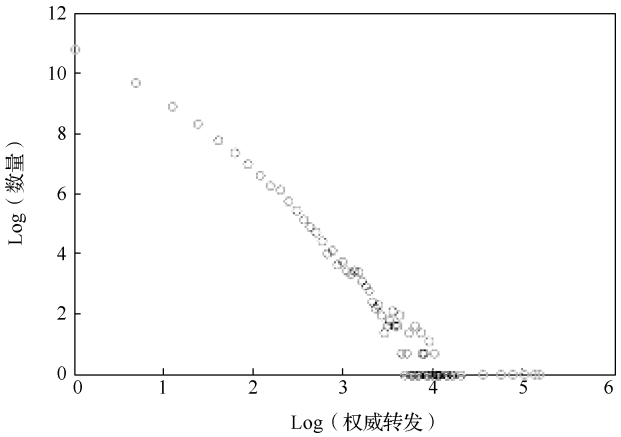


图 3-4　微博转发关注边数分布

（1）权威比率、微网络结构对转发影响的权重。从表 3-4 可以看出，在 Pattern I 中，权威比率为 Small 的权重比较大，表明权威比率影响较小，说明在 Pattern I 中包含了大量的强连接关系，社会地位相当的用户所占比例更大一些。在 Pattern II 中，权威比率为 Medium 的权重比较大，表明权威比率影响较小，说明 Pattern II 中包含了大量的弱连接关系。在 Pattern III 中，权威比率为 Large 的权重最大，说明在 Pattern III 中包含了更多的权威连接关系。

（2）地理距离、微网络结构对转发影响的权重。从表 3-5 可以看出，在 Pattern I 中，地理距离为 Near 的权重比较大；在 Pattern II 中，地理距离为 Far 的权重比较大；在 Pattern III 中，地理距离为 Medium 的权重比较大。这说明 Pattern I 和 Pattern II 容易受到地理距离的

影响，Pattern I 中的用户之间地理距离更近，而 Pattern II 中的用户之间地理距离更远。这是因为在强连接关系中，用户为了维持这种关系需要花费大量的时间、情感等，而近距离比远距离更容易维持这种关系。在弱连接关系中，远距离的用户反而更容易被关注，其原因是远距离的用户更多来自不同的社区或团体，可能包含更多的新颖的信息，更容易受到用户的关注。在 Pattern III 中，大多数是权威连接关系，一般用户只需通过公共媒体，例如电视、新闻等，单方面地了解权威人士，因此几乎不受地理距离的限制。

表 3-4　权威比率、微网络结构对转发影响的权重

		微网络结构		
		Pattern I	Pattern II	Pattern III
权威比率	Small	0.078	0.063	–0.141
	Medium	–0.114	0.119	–0.005
	Large	0.036	–0.182	0.146

表 3-5　地理距离、微网络结构对转发影响的权重

		微网络结构		
		Pattern I	Pattern II	Pattern III
地理距离	Near	0.133	–0.128	–0.005
	Medium	–0.045	–0.051	0.096
	Far	–0.088	0.179	–0.091

（3）性别、微网络结构对转发影响的权重。从表 3-6 可以看出，在 Pattern I 中，无论男性用户还是女性用户的微博，都会受到女性用户的关注（FM，FF），这是因为在 Pattern I 中包含了大量的强连接关系，用户之间大多是朋友、亲戚关系，女性用户更愿意与同性朋友交往。在 Pattern II 中，无论男性用户还是女性用户的微博，都会受到男性用户的关注（MM，MF），这是因为在 Pattern II 中包含了大量的弱连接关系，用户一般来自不同的社区或团体，能够提供新颖的信息，男性用户比较关注新颖的信息。在 Pattern III 中，无论男性用户还是女性用户的微博，都会受到女性用户的关注（FM，FF），说明服从权威因素对女性的影响更大一些。

表 3-6　性别、微网络结构对转发影响的权重

		微网络结构		
		Pattern I	Pattern II	Pattern III
性　别	MM	–0.080	0.111	–0.031
	MF	–0.141	0.215	–0.074
	FM	0.043	–0.062	0.019
	FF	0.178	–0.264	0.096

（4）权威比率对转发影响的权重。随着权威比率的增大，对微博转发影响的比重也在

增大，这意味着权威比率增大导致了微博更容易转发，表明权威比率对转发行为有着较大的影响，权威比率与转发行为之间有着很强的关联性。可以解释为随着权威比率的增大，社会权威差距在增大，服从权威效应逐渐显露出来，导致用户容易转发权威人士的信息。因此，服从权威对转发行为有着显著地影响。

（5）微网络结构对转发影响的权重。在 Pattern I 中，容易发生转发行为，这是因为 Pattern I 包含了强连接关系，两个用户大多来自于相同的社区或团体，有着共同的价值观和认同感，因此更加容易接受对方的信息。相反地，在 Pattern II 中，不容易发生转发行为，这是因为 Pattern II 中包含了弱连接关系，用户可能来自于不同的社区或团体，他们的价值观也不一样，因此较难接受对方的信息，除非是新颖的信息。在 Pattern III 中，与 Pattern I 中用户较容易转发和 Pattern II 中用户较难转发相比，Pattern III 中的用户转发行为处于中间水平，表明用户对权威连接关系的微博转发意愿处于强连接与弱连接中间。

综上所述，微网络结构、权威比率、地理距离及性别 4 个属性对微博转发行为具有不同的影响：

（1）不同的微网络结构对微博转发行为的影响不同。Pattern I 中包含了较多的强连接关系，用户一般来自于相同的社区或团体，有着共同的价值观和认同感，更加容易接受对方的信息，用户之间更容易产生微博转发行为。Pattern II 中包含了较多的弱连接关系，用户一般来自于不同的社区或团体，他们的价值观也不一样，接受对方的信息比较困难，用户之间不容易产生微博转发行为。Pattern III 中包含了权威连接关系，用户对权威连接关系的微博转发意愿处于前两者之间。

（2）权威比率对微博转发行为的影响因微网络结构而异。权威比率对 Pattern I 中的微博转发行为影响较小，这是因为 Pattern I 中包含了较多的强连接关系，微博转发比较容易，而权威比率的增大并不会对微博转发行为产生更大的影响。权威比率的增大将使得 Pattern II 中的微博转发更加困难，这是因为来自于不同社区或团体的用户对权威人士或名人的抵触心理比较强烈，很难转发他们的微博。权威比率对 Pattern III 中的微博转发行为影响最大，随着权威比率增大，微博转发更加容易，服从权威效应逐渐显露出来，用户容易转发权威人士的信息。

（3）地理距离对微博转发行为的影响与微网络结构有关。在 Pattern I 中，近距离用户的微博更容易转发，因为较近的地理距离分布着更多的强连接关系。在 Pattern II 中，远距离用户的微博更容易转发，因为远距离的用户微博包含更多的新颖信息，用户对新颖的微博怀有更多的好奇心。在 Pattern III 中，不同地理距离对转发影响不大，这表明权威连接关系的建立不受用户之间地理距离的影响。

（4）性别对微博转发行为的影响与权威比率有关。女性用户比较关注具有强连接关系的朋友微博，更愿意与同性朋友交往。男性用户比较关注来自不同社区或团体的新颖信息，更关注女性的微博。服从权威因素对女性的影响更大一些，说明女性更崇拜网络名人。

3.3 微博转发行为预测

随着 Web 2.0 的发展，微博网络已经成为互联网中最流行的信息共享和分发平台。微博转发是微博网络中最重要的信息传播机制，如果能够准确地预测微博用户转发行为，那么就能够预测该微博的传播方向、次数以及覆盖范围等，对于网络舆情发现、突发事件预测以及用户推荐等方面具有重要的现实意义，因此越来越受到研究者的重视。

下面介绍一种基于社会纽带关系的微博转发预测方法，首先基于社会纽带关系提取用户之间的微网络结构、权威比率以及其他特征，然后根据这些特征对微博转发的权重大小，采用基于随机森林的预测算法对微博转发行为进行预测。

3.3.1 预测算法

随机森林是一种分类预测算法，由于决策树是随机森林的基本分类器，因此首先介绍决策树及相关概念，然后引出随机森林，并对随机森林泛化误差的内部估计、效能强度、相关系数以及参数优化等问题进行分析。

1．决策树算法

决策树是一种类似流程图的树结构，由节点和有向边组成。树中包括三类节点：根节点、内部节点和叶子节点。其中，根节点位于树的最顶层；内部节点代表一个属性分裂问题，每个分裂输出代表一个分枝；叶子节点是终节点，存放着带分类标签的数据集。从根节点到叶子节点的每一条路径都形成一个分类。决策树的算法很多，如 ID3、C4.5、CART 等，这些算法通常采用自上而下的贪婪算法来构造，通用的决策树算法步骤如下。

算法 3-1　通用决策树算法

输入：训练数据集 D
　　候选属性的集合 attribute_list
　　划分准则 Attribute_selection_method，由分裂属性和分裂点组成
输出：一棵决策树
（1）创建一个节点 N;
（2）If D 中元组都是同一类 C then
（3）　　返回 N 为叶子节点，以类 C 标记；
（4）If attribute_list 为空 then
（5）　　返回 N 为叶子节点，标记为 D 中的多数类；
（6）使用 Attribute_selection_method（D, attribute_list），找出最好 splitting_criterion；
（7）用 splitting_criterion 标记节点 N；
（8）If splitting_attribute 是离散的并且允许多路划分 then
（9）　　从 attribute_list 中删除属性 splitting_attribute；
（10）For splitting_criterion 的每个输出 j

```
（11）　设 Dj 是 D 中满足输出 j 的数据元组的集合；
（12）　If Dj 为空 then
（13）　　加一个树叶到节点 N，标记为 D 中的多数类；
（14）　Else
（15）　　加一个由 Generate_decision_tree（Dj, attribute_list）返回的节点到节点 N;
（16）End for
（17）返回 N;
```

该算法需要输入 3 个参数：D、attribute_list、Attribute_selection_method。其中，D 是训练数据集，attribute_list 是描述数据元组的属性列表，Attribute_selection_method 是指定选择属性的启发式方法，该方法通常使用一种属性选择度量，其中包括信息增益、信息增益比率以及基尼（Gini）指数等。通用的决策树算法过程都基本类似，而各个算法的差别在于在创建树时属性选择方法以及剪枝方法。

在决策树创建时，由于数据中的噪声和离群点，许多分枝反映的是训练数据的异常，需要剪去这些不可靠的分枝。剪枝方法用来处理这种过分拟合问题，通常有先剪枝和后剪枝两种方法。

在先剪枝方法中，通过提前停止树的构造来实现对树的剪枝。例如，在构造树的时候，可以使用诸如统计显著性、信息增益、Gini 指数等度量来评估分裂的优劣。如果划分一个节点的属性值低于预定义的阈值，则停止分裂。

后剪枝方法中，首先让树充分生长之后，再判断是否剪去一些分枝。常用的方法包括根据错误分类率对决策树进行事后修剪等。

在决策树学习过程中，虽然剪枝可以减少过度拟合问题，但是剪枝方法以及相应的度量选取最终影响决策树的优劣，而这一选取过程是决策树学习过程中的一个难点。另外，单棵决策树的准确度也容易受到训练数据的影响而导致分类结果的不稳定。为了克服这些缺点，引入一个新的预测模型，即随机森林模型。

2．随机森林算法

单棵决策树对数据分类有较好的准确度，然而这种分类准确度易受到训练数据本身的影响，具有不稳定性。为了避免这种不稳定性，一种解决方案是产生多棵决策树，参与投票，选出最好的分类，这就是随机森林的思想。

随机森林就是采用随机的方式建立一个森林，森林里面包括多棵决策树，且每一棵决策树之间没有关联。在建立森林后，当有新的输入样本数据进入时，就让每棵决策树进行判断分类，当所有树判断分类结束，组合多棵树的预测，通过某种投票方式得到最终预测结果。

1）算法构造

随机森林定义为由一组决策树分类器 $(h(X,\theta_k),k=1,\cdots,K)$ 组成的集成分类器，其中 $\{\theta_k\}$ 是服从独立同分布的随机变量，K 表示随机森林中决策树的个数，在给定自变量 X 下，每个决策树分类器通过投票来决定最优的分类结果。

随机森林算法的一般构造过程如下。

（1）给定全部训练数据，随机抽取部分数据，形成新的子样本数据；

（2）对新的子样本数据中 M 个特征变量，随机抽取 m（$m<M$）个特征，然后构造完整的决策树；

（3）重复步骤（1）、（2），得到 K 个决策树，形成随机森林；

（4）每个决策树参与投票，最终以某种投票方式，选出最优的分类。

图 3-5 给出了通用随机森林构造过程图，从图 3-5 可以看出，所有决策树分类器都是并行的，也就是说，每个决策树分类器的构造过程互不影响，相互独立。

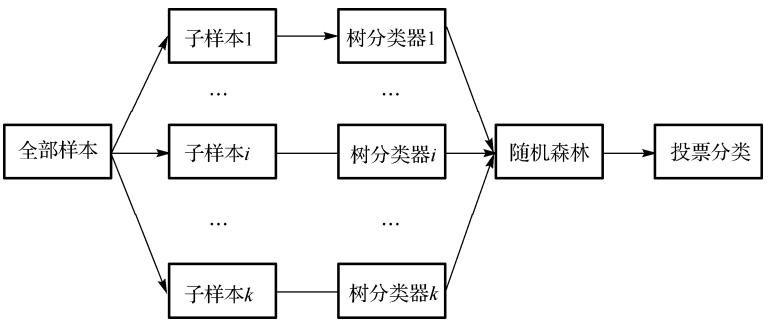


图 3-5　随机森林构建过程图

对于子样本采集过程，随机森林采用了两个随机采样方法，保证了后续的决策树分类器之间的独立性。

首先，该算法对输入的数据进行采样，即行采样。在此过程中，一个常用的方法是 bootstrap 方法，它是一种有放回的抽样方法，也就是在抽样得到的样本集合中，允许重复抽样。假设输入样本为 N 个，那么重复抽样的样本也为 N 个，这样使得在训练时抽取的子样本并不是全部的样本，当 N 值足够大时，约 63.2%的输入样本成为子样本。数据采样还有其他采样方法，如保持、K 折交叉验证等方法。

其次，该算法对采样子数据的特征再次进行采样，即列采样。该过程通常从 M 个特征中以某种随机方法选取 m 个（$m \ll M$）特征，关于 m 的取值下面将详细地介绍。列采样通常有两种方法，第一种方法是随机选择特征，每一个节点随机选取一小组输入变量进行分割，节点分割是根据 F 个选定特征，而不是所有的特征来决定，然后利用 CART（Classification And Regression Tree）方法完全生成决策树，一旦决策树完全构建完成，使用多数表决法来预测，在构建过程中选择的输入变量 F 是固定的，F 的大小影响随机森林的强度和相关性，如果 F 足够小，树的相关性减弱，同时分类模型的强度也在减弱。由于每个节点只需要考察输入变量的一个子集，该方法的运行时间较少；第二种方法是随机选取特征变量的线性组合，随机选择 L 个输入变量进行线性组合得到新的特征，在每个节点上，随机选出 L 个变量 v_1,

v_2, …, v_L 以及相应的 L 个随机数 k_1, k_2, …, k_L，且每个随机数 $k_i \in [-1, 1]$，做线性组合 $V = \sum_{i=1}^{L} k_i v_i$ 。

与一般决策树构造不同的是，随机森林决策树不需要剪枝，即所有的决策树都是完全分裂的。在这个过程中，所有决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本都是指向同一个分类。一般通用的决策树算法都需要进行剪枝，但是随机森林并不需要这样做，一个重要的原因是两个随机的采样过程保证了数据的随机性和独立性，不会出现 over-fitting 现象。

另外，不同随机森林的投票机制也不同，当一个被预测实例进入时，每个决策树分类器都要进行预测分类，然后以某种方式参与投票。因此，不同的投票机制将导致不同的分类结果。投票机制主要分为两大类：简单投票机制和贝叶斯投票机制。

简单投票机制的基本思想是多个基本分类器都进行分类预测，然后根据分类结果用一种投票方法进行表决。投票方法可以分为一票否决、少数服从多数和阈值表决等方法。一票否决法是指预测结果当且仅当所有决策树分类器都把预测的实例划分到某一类时才有效，否则拒绝这个实例分类。少数服从多数法是指每个决策树分类器预测分类，然后进行投票，得票数多的那个类作为实例的最终分类结果。阈值表决法是指统计实例被决策树分类器划分和不划分某类的得票数，当两者的比例超过预定义的阈值时，就划分到此类中。

贝叶斯投票机制不同于简单投票机制，简单投票机制假设每个决策树都是平等的，没有分类能力的差别，但是这种假设并不总是合适的。例如，在实际生活中，听取一个人的意见时会考虑这个人过去的意见是否有用。贝叶斯投票方法就是基于这种思想提出的，贝叶斯投票方法是基于每个决策树在过去分类的表现来设定一个权值的，然后按照这个权值进行投票，其中每个决策树权值可以利用贝叶斯定理计算出来。

理论上，贝叶斯投票方法在假设空间中所有假设的先验概率都正确的情况下能够获得很好的效果，但是在实际应用中往往不可能穷尽整个假设空间，也不可能准确地给每个假设分配先验概率。因此，在实际使用过程中，简单投票方法优于贝叶斯投票方法。

2）参数优化

决策树分类器个数 K 以及特征个数 m 的选取也直接影响着随机森林的性能，下面讨论如何选取最优参数 K 和 m。

随着决策树分类器个数 K 的增加，随机森林不会出现过度拟合问题，但是会产生一个有限的泛化误差。分类器泛化误差定义为：

$$\mathrm{PE} = P_{X,Y}(mg(X,Y) < 0) \tag{3-3}$$

式中，PE 为泛化误差，$P_{X,Y}$ 的下标 X,Y 表示概率覆盖的定义空间。

分类器泛化误差的上界为：

$$PE \leqslant \frac{\bar{\rho}(1-s^2)}{s^2} \tag{3-4}$$

式中，$\bar{\rho}$ 为分类器相关性均值，s 为分类器效能强度。

可见，随机森林的泛化误差上界包含两个要素：决策树的分类效能强度 $\bar{\rho}$ 和决策树间的相关性 s，可以由 $\bar{\rho}$ 和 s 两个参数推导出来，其中与 $\bar{\rho}$ 成正比，与 s^2 成反比，因此比值 $\bar{\rho}/s^2$ 越小越好。对于多于两个分类的情况，效能强度依赖于随机森林以及相应的每个决策树。

随着特征个数 m 的增加，分类器的相关性 $\bar{\rho}$ 和效能强度 s^2 也相应地增加，反之亦然。因此，m 值会产生一定的泛化误差，当 m 值在某一区间时，泛化误差上界将处于最低。

3）特征提取

在分类预测中，一个重要的任务是寻找相关的重要特征。一方面在数据集中有许多特征与分类预测无关，另一方面有些特征可能是冗余的。如果选择的特征不具有辨别能力，则会直接影响到分类器性能。如果选择了具有充分辨别能力的特征，则会极大地提高分类器的预测精度。因此，特征选取过程是至关重要的。

微博转发行为在社交网络信息传播中有两个重要过程：同化与社会影响，同化过程导致了用户之间网络结构的变化，而社会影响过程则导致了两个用户的属性变化。用户之间不同类型的网络结构以及属性关系代表着不同类型的用户关系，这意味着微博是否会存在转发行为，这些特征对预测微博转发行为具有重要的作用。因此，需要从网络结构和用户属性中提取特征，包括权威比率、微网络结构、地理距离、用户性别以及用户自身属性值等。关于权威比率、微网络结构、地理距离、用户性别等特征见 3.2.1 节，下面是用户属性特征的提取。

对于关注边 A→B，用户 A、B 都有基本资料，可以从中提取用户的自身属性特征，其中包括 9 个特征，如表 3-7 所示。用户 A、B 都有自身属性特征，因此共提取了 22 个相关特征。

表 3-7　用户自身属性特征

特　征	描　述
1	居住省份
2	居住城市
3	关注人数
4	粉丝数
5	微博数
6	喜欢的人数目
7	注册时间
8	访问权限
9	是否为认证用户

4）算法实现

在随机森林预测算法中，一个重要的问题是如何进行随机数据采样。构造一组相互独立的决策树，数据采样包括数据集采样和数据特征采样。

对于数据集随机采样，经典的随机森林方法采用了 bootstrap 方法，即从数据集中有放回均匀抽样。在平均情况下，63.2%的数据集将作为样本子集训练模型，其余 36.8%数据集作为袋外测试数据集检验模型。对于小数据集，bootstrap 方法效果很好。随着数据集增大，bootstrap 方法效果并不理想。由于微博数据集比较大，因此采用 K 折交叉验证方法来随机采集微博数据集。在该方法中，数据集被随机地采集，并划分为 K 个互不相交且大小一致的子集，然后利用数据对算法进行 K 次训练与测试，其中在第 i 次时，第 i 个子集作为测试数据，而其余的所有数据子集一起作为训练数据。通常实验将采用 10 折交叉验证，这是因为它具有相对较低的偏倚和方差。

对于数据特征的采样，经典的随机森林算法采用完全随机方法，即所有特征被抽取到的概率都相同。这里采用一种基于特征权重的方法来抽取特征，微博转发影响权重越大的特征越容易被抽取到，这是因为权重越大的特征对微博转发预测作用也越大，这样可以提高预测模型的准确度。其中，特征权重采用信息增益算法来刻画，通过特征的信息增益值来代表其权重大少，当特征的信息增益值越大，则该特征的影响权重也越大。

对于数据特征的采样，一方面，算法首先利用信息增益算法计算出所有特征的信息增益值并进行排序，然后根据特征的信息增益值去除对微博转发影响权重很小的特征，如果选取了那些权重很小的特征，则生成的决策树分类器区分度很弱，反而导致预测误差增大。另一方面，算法将根据不同特征的权重来选取特征，对微博转发影响权重越大的特征，被抽取到的概率也越大，有利于提高决策树分类器的准确度。

对于决策树的特征属性度量方法，采用 Gini 指数，即 CART 分类回归树方法。这是基于两方面考虑的，一方面是部分选取的特征是连续的，CART 方法能够处理这类特征；另一方面是微博所提取的数据特征中类的数量较少，CART 方法能够很好地处理这类数据，不会因类的数量太少而产生度量偏差的问题。

对于分类预测的投票机制，采用简单多数投票法，即数目最多的类就是最终的类，分类决策公式如下：

$$H(x)=\arg\lim_{Y}\sum_{i}I(h_i(x)=Y) \tag{3-5}$$

式中，$H(x)$表示组合分类模型，$h(x)$表示单个决策树分类模型。

一种改进的随机森林微博转发预测算法（IRFMR）具体步骤如下。

算法 3-2　随机森林微博转发预测算法（IRFMR）

输入：微博数据集 S

微博预测数据集 P

模型训练：

（1）对数据集 S 用 10 折交叉验证方法采样，得到新的训练数据集 S_n。

（2）对数据集 S_n，计算信息增益算法计算每个特征的权重，排序并排除小于设定阈值的特征。

（3）对于训练集 S_n 所有大于设定阈值的 M 个特征，基于特征的权重大小，随机选取 m（$m<<M$）个特征，构成新的数据集 S_m。

（4）对数据集构造完整的决策树（CART 方法），不进行剪枝。

（5）循环步骤（1）、（2）、（3）、（4），直到 K 个决策树建立，随机森林构造完成。

预测：

（6）对数据集 P 的一个变量 x 分类标签，每棵决策树进行投票。

（7）计算所有投票数 $H(x)$，票数最高的分类就是变量 x 的分类标签。

（8）循环（6）、（7），直到数据集 P 所有变量的分类标签被标记。

输出：预测数据集 P 的分类标签。

3.3.2 算法验证

下面通过实验数据对微博转发行为预测算法性能进行测试和验证。

1．实验数据集

实验数据来源于新浪微博。从 2011 年 5 月至 2011 年 7 月，随机采集了 171 769 个用户以及相应的 702 789 条活跃的关注边，其中用户包括标签、个人资料等信息，关注边则包括两个用户转发关系。如果一条关注边中的两个用户存在转发行为，则该关注边为正例，否则为负例，最终得到 185 237 个正例和 517 552 个负例。

Weka 作为一个公共的数据挖掘与分析平台，集合了大量数据挖掘算法，包括了数据预处理、分类、回归、聚类、关联规则等，因此所有实验数据都在 Weka 平台上运行。

2．参数优化

算法对全部数据进行采样，有部分数据不在训练样本中，这些数据称为袋外数据（OOB），使用袋外数据测试模型性能称为 OOB 误差估计。在随机森林算法中，对每一棵决策树测试，可以得到一个 OOB 误差估计。对所有决策树的 OOB 误差估计取平均值，即可得到随机森林的泛化误差估计，实验证明，随机森林的 OOB 误差估计是无偏估计。

用 OOB 误差估计来选取决策树的个数 K 以及特征的个数 m，当 OOB 估计最少时，则参数 K 和 m 为最优。由于 K 和 m 两个参数都对 OOB 误差估计有影响，两个参数组合，需要进行 $K\times m$ 次 OOB 误差估计比较。为了简化实验过程，在估计一个参数时需要固定另一个参数值，这样只需比较 $K+m$ 次。另外，由于数据集较大，每次计算 OOB 误差估计值时间较长，抽样 5%数据集进行试验。

图 3-6 给出了当 $m=2$ 时决策树个数 K 与 OOB 误差估计值的曲线图，随着 K 值的增加，OOB 误差估计值在减少，但是下降趋势不同。当 K 值处于 5～17 区间时，开始阶段 OOB 误差估计值下降较快，随后逐步地减弱。当 K 处于 17～20 区间时，OOB 误差估计值下降趋势明显地减弱，已趋于平稳，这说明 OOB 误差估计接近收敛。

图 3-7 给出了算法运行时间与 K 值的关系，从图 3-7 可以看出，随着 K 值增大，算法运行时间在逐步增加，综合考虑 OOB 误差估计及算法运行时间因素，则 $K=15$ 为最优。

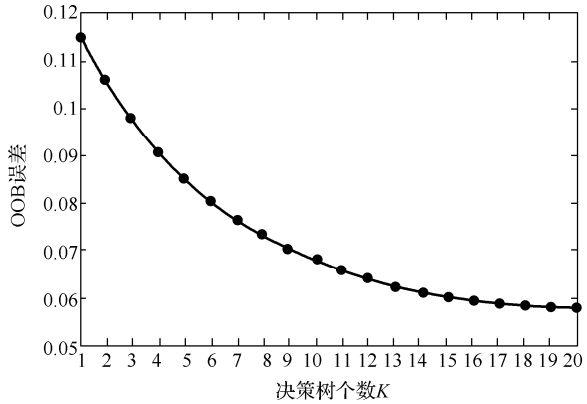


图 3-6　决策树个数 K 与 OOB 误差估计值曲线

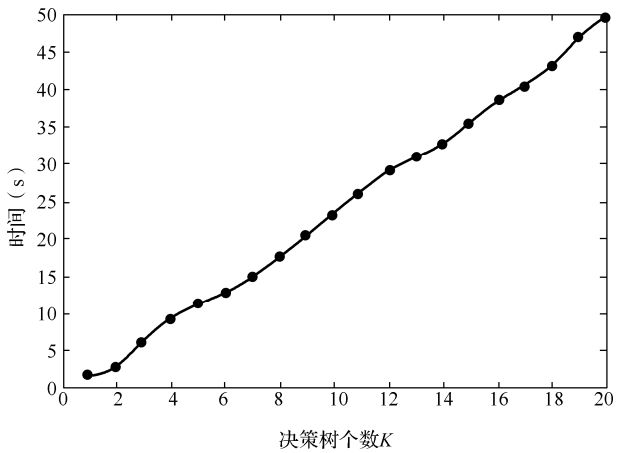


图 3-7　算法运行时间与 K 值的关系

图 3-8 给出了当 $K=15$ 时特征个数 m 与 OOB 误差估计的曲线变化，从图 3-8 可以看出，随着 m 值增加，OOB 误差估计值先迅速地下降，然后逐步地回升，当 $m=3$ 时，它处于最小值，实验结果最优。因此，将 IRFMR 算法的参数设定为 $K=15$，$m=3$。

3．算法性能验证

下面将 IRFMR 算法与逻辑回归（LR）、决策树（DT）、Adaboost（Ada）、朴素贝叶斯（NB）、多层感知器（MP）及经典随机森林方法（RF）等几种经典的分类算法进行对比实验。所有的算法都是在参数最优情况下得到的结果。例如，在经典随机森林算法中，当决策

树个数 $K=15$ 和特征个数 $m=3$ 时，该算法性能最优，选取此时的结果作比较。评价指标为准确率（P）、召回率（R）、F_1 度量（F_1）、ROC（Receiver Operating Characteristic），其中，准确率是指算法预测出的用户关系数量与数据集中所有用户关系数量之比；准确率也称为查准率，其计算公式为 $P=A/(A+B)$，A 为正确预测的数量，B 为错误预测的数量；召回率也称为查全率，其计算公式为 $R=A/(A+C)$，A 为正确预测的数量，C 为未预测出的数量；F_1 度量是为了平衡准确率和召回率，其计算公式为 $F_1=2PR/(P+R)$，F_1 值越高，其综合性能越好；ROC 也称为受试者工作特性曲线，在机器学习和数据挖掘领域中应用时，主要用于度量和评价分类器算法的性能，ROC 值越高，其分类器算法性能越好。

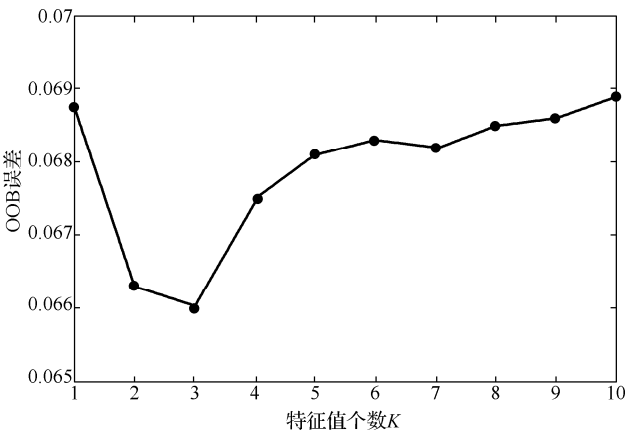


图 3-8　OOB 误差估计与特征值个数 m 变化曲线

表 3-8 给出了正例预测中各个算法比较的结果，在正例预测中，不同算法效果相差明显。IRFMR 算法在各项指标上都明显优于其他算法，其中准确率为 94.8%，召回率为 70.3%，F_1 值为 80.7%，ROC 值为 94.9%，其次是 RF 算法，NB 算法表现最差。在召回率指标上，各个算法都不高，其原因是由于正负例的比例不均衡，大量被错误分类的负例个数降低了召回率。

表 3-8　正例预测中各个算法性能对比

	准　确　率	召　回　率	F_1 值	ROC
LR	0.645	0.133	0.221	0.790
NB	0.359	0.282	0.316	0.740
DT	0.714	0.629	0.669	0.852
MP	0.709	0.254	0.374	0.816
Ada	0.615	0.142	0.231	0.794
RF	0.929	0.681	0.785	0.917
IRFMF	0.948	0.703	0.807	0.949

表 3-9 给出了负例预测中各个算法比较的结果，相比于正例，各个算法的各项指标都较高，准确率和召回率都接近或超过 90%，说明所有算法在预测负例时效果都比较理想。而 IRFMR 算法在各项指标上都优于其他算法，说明在负例预测中，IRFMR 算法性能也优于其他算法。

表 3-9　负例中各个算法性能对比

	准　确　率	召　回　率	F_1 值	ROC
LR	0.883	0.989	0.933	0.790
NB	0.895	0.924	0.909	0.740
DT	0.945	0.962	0.953	0.852
MP	0.897	0.984	0.939	0.816
Ada	0.883	0.986	0.932	0.794
RF	0.953	0.992	0.977	0.917
IRFMF	0.957	0.994	0.979	0.949

综合所有的正负例预测，IRFMR 算法的准确率高达 95%，表明该算法可以成功地预测 95%的微博转发。因此，IRFMR 算法能够很好地预测用户间微博转发行为。

4．特征分析

采用信息增益算法来分析所提取的特征对微博转发预测的作用，特征的信息增益值越大，则该特征在微博转发预测的作用也就越大。图 3-9 给出 22 个特征值的权重值，其中 NS、GE、PR、LC 分别表示用户间的微网络结构、性别关系、权威比率、地理位置，A1～A9 分别表示用户 A 的 9 个自身属性特征，而 B1～B9 表示用户 B 的 9 个属性特征，与用户 A 属性特征相同，用户自身属性特征如表 3-7 所示。

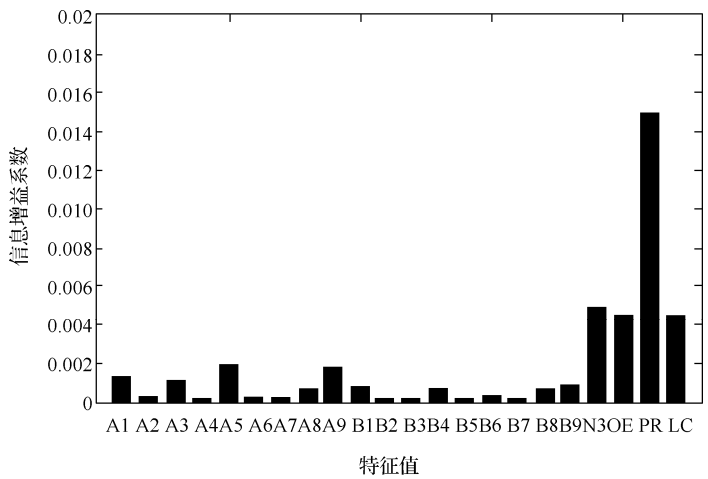


图 3-9　22 个特征值的权重值

从图 3-9 可以看出，权威比率的系数最显著，是其他特征值的 3 倍以上，这说明权威比率在预测微博转发的作用最为显著，其次分别是微网络结构、性别关系以及地理位置。而用户 A、B 自身属性特征的系数总体上偏弱，尤其一些特征系数已经接近零值，这说明用户自身属性特征对微博预测的作用较弱。从以上分析可知，GE、LC、PR 和 NS 是微博转发预测中最重要的 4 个特征。

综上所述，在微博网络中，权威比率、微网络结构、性别关系及地理位置特征对微博用户转发行为产生重要的影响，通过提取这些特征，并采用 IRFMR 算法对微博用户转发行为进行预测，能够较好地预测微博用户转发行为。

3.4 微博转发峰值分析

微博具有很强的时效性，不同时刻的微博关注度是不同的，关注度的时间序列反映了微博受欢迎程度的变化，在某一时刻的微博关注度达到峰值则表明了此时的微博最受用户欢迎和关注。因此，时间序列峰值是微博转发时间序列的最重要特征，对于微博舆情监测具有重要的意义。

3.4.1 时间序列概念

微博转发时间序列是指微博转发次数随着时间变化的曲线，下面给出相关时间序列的概念和定义。

给定 t_n 为微博发布后的第 n 个时间间隔，x_{t_n} 是在 t_n 时段内的微博转发数，则微博转发时间序列定义为：

$$X=\{x_{t_1},x_{t_2},\cdots,x_{t_n}\} \tag{3-6}$$

假定阈值 n_p 是微博转发时间序列峰值的临界点，则在峰值 X_p 中有：

$$x_{t_k}\begin{cases}\geqslant n_p & t_k\in[t_i,t_j]\\ <n_p & t_k\in(t_{i-1},t_{j+1})\end{cases} \tag{3-7}$$

式中，x_{t_k} 表示在峰值 X_p 内的转发次数，t_i、t_j 分别为峰值 X_p 的开始时间和结束时间。峰值 X_p 的时段长度 T_p 定义为：

$$T=t_j-t_{i-1} \tag{3-8}$$

假定 T_{p_i} 是微博转发时间序列的第 i 个峰值的时段长度，则总峰值时间定义为：

$$T_{\text{pall}}=\sum_i T_{p_i} \tag{3-9}$$

总峰值时间是微博转发时间序列所有峰值时段之和。

给定 $t_0 = 0$ 是原始微博的发布时间，t_e 是微博转发时间序列最后一个峰值的结束时间，则微博存活时间定义为：

$$T_d = t_e - t_0 = t_e \tag{3-10}$$

微博存活时间是微博整个生存周期，其中包括了峰值时段和非峰值时段。

图 3-10 为微博转发时间序列示例图，该微博转发时间序列存在 3 个峰值，其中每个峰值幅度以及时段长度都不一样。总峰值时间是 3 个峰值时段长度之和，而存活时间是指微博发布时间开始到最后一个峰值结束这一时间段。

由此可见，总峰值时间与存活时间并不一样。总峰值时间主要用来描述微博欢迎度最高的时间段长短，而存活时间则描述微博的存活时间长短。

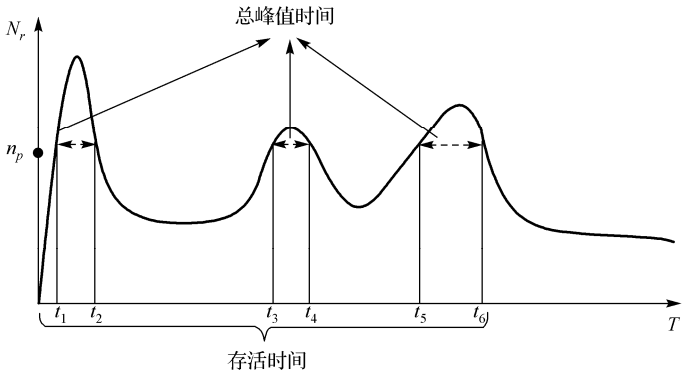


图 3-10　微博转发时间序列示例

给定 $A \rightarrow B \rightarrow \cdots J \rightarrow K$ 是一条微博转发路径，N_{re} 是该转发路径上用户转发次数，则微博转发路径长度定义为：

$$P_l = N_{re} \tag{3-11}$$

微博转发路径长度反映了微博传播的深度，是从网络结构上来研究微博转发的性质。

3.4.2　峰值特性分析

下面通过实验数据对微博转发峰值特性进行分析。

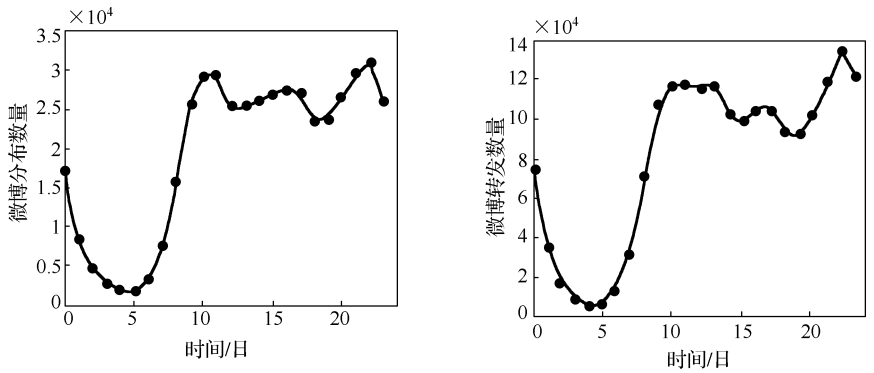
1．实验数据集

实验数据来源于新浪微博，数据集中包含了微博转发记录，微博转发记录包括了 5 636 858 个用户发布的 46 584 914 条微博以及相应的 190 920 026 条转发记录。

2．微博时间特性

在微博发布或转发过程中存在着周期性。例如，在一天中的不同时间段用户发布或转

发微博数目是不同的。以 24 小时作为周期，图 3-11 给出了不同时段微博发布和转发次数的分布，微博发布和转发次数的曲线变化大体上相同，且与人们一天的作息规律相一致。例如，从早上 8 点开始，用户开始活动，发布和转发微博次数逐渐增加，并一直持续在较高的区间，直到凌晨零点后才下降。在这一区间出现了两个峰值，第一个峰值出现在上午 10 点左右，持续时间达 2 小时；第二个峰值出现在晚上 9 点左右，持续时间达 3 小时，也是一天中最高的峰值，表明在这两个时间段微博用户最多，尤其是晚上时间段。在中午 12 点和下午 6 点左右，曲线都会出现一个小波谷，表明在这两个时间段用户在用餐或休息。从凌晨 0 点开始至早上 8 点前，微博发布和转发数是一天中最少的，其中最低谷出现在凌晨 5 点，不到最大峰值的 1/20。



（a）用户发布微博次数　　（b）用户转发微博次数

图 3-11　24 小时微博发布和转发次数分布图

另一个问题是微博的时效性，微博的时效性可以看作用户对微博关注度随着时间的变化过程，微博关注度可以用微博转发次数来表征，例如某一时刻微博转发次数越多，表明此时微博的关注度也越高。因此，微博的时效性主要是研究微博转发次数的时间序列变化。图 3-12 给出了微博转发时间间隔分布，符合幂律分布，大部分微博的转发时间间隔较短，具有很强的时效性，只有小部分微博的转发时间间隔较长。例如，81.8%的微博在一天内被转发，而转发时间超过两天的微博数目不足 12%，并且随着时间推移，其微博数量呈指数下降。这说明微博具有很强的时效性。

3．微博转发峰值检测

为了更好地分析时间序列峰值，从数据集中检测出转发次数超过 100 的微博峰值分布，总共得到 207 259 条微博及相应的 48 302 776 次转发记录，平均每条微博转发次数约为 240 次，然后将这些微博转化为相应的转发时间序列，每一个时间刻度为 6 小时。

表 3-10 给出了微博峰值检测结果，其中没有峰值的微博有 26 620 条，大约占 12.84%，说明微博网络存在大量不活跃的微博；含有一个峰值的微博有 142 406 条，大约占

68.71%，属于正常的微博；含有多个峰值的微博有 38 233 条，大约占 18.45%，有些微博的峰值甚至达到 12 个，说明微博容易受到其他因素影响而导致多次被高度关注。

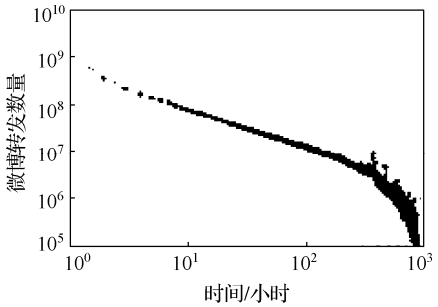


图 3-12　微博转发时间间隔分布

表 3-10　不同峰值的微博数量及比例

峰值个数	微博数目	比　例
0	26 620	12.84%
1	142 406	68.71%
>=2	38 233	18.45%

表 3-11 给出了不同峰值的微博的平均转发路径长度，其中峰值为 0 和 1 的微博的平均路径长度比较接近，说明这两类微博的传播深度相类似。多个峰值的微博的平均路径长度达到了 2.32，说明多个峰值的微博的转发路径较长，传播深度较深。

表 3-11　不同峰值的微博平均转发路径长度

峰值个数	平均转发路径长度
0	2.03
1	2.09
>=2	2.32

4．微博转发峰值分析

微博转发峰值分析主要针对不同类型热门主题和用户的微博，分析这些微博的各个特征，包括总峰值时间、存活时间、转发路径长度以及峰值时段路径长度等。

1）微博主题类型分析

微博中包括了 44 个不同的热门主题，这里提取了转发次数前 4 位的主题微博进行分析，它们是房价问题事件、河北大学校园飙车致死案事件、李阳家暴事件以及小米手机发布事件。另外，对于不含任何热门主题的微博，也看作一种类型的微博。

表 3-12 给出了不同类型主题微博的特征比较，它们有两个特点：一是含有热门主题和不含热门主题微博特征存在较大的差异，不含热门主题微博的转发路径长度（P_i）、总峰值

时间（T_p）及存活时间（T_d）等都较短，转发路径长度只有 1.90，大约为含有热门主题微博的转发路径长度的 60%～70%，这说明不含热门主题的微博被关注时间和存活时间都较短，传播范围有限；二是不同类型热门主题的微博特征也不同，例如，河北大学飙车致死事件的转发路径和存活时间较长，而小米手机发布事件的转发路径和存活时间则较短，这说明河北大学飙车致死事件的社会影响大，受到用户的高度关注，具有较长的存活时间。

表 3-12　不同类型主题微博比较

主题类型	Average(P_l)	Average(T_p)	Average(T_d)(days)
不含任何热门主题	1.90	2.95	15.85
房价问题	2.66	3.05	22.75
小米手机发布	2.62	3.39	16.32
李阳家暴事件	2.89	3.71	21.20
河北大学飙车致死事件	2.97	3.64	28.15

图 3-13 给出了不同类型主题微博的转发路径分布图，随着转发路径长度增加，不同主题微博所占比率都在下降，但是下降幅度有所不同。在不含热门主题微博中，转发路径较短的微博比率较大，其中路径长度小于 3 的微博比率接近 84%；而路径长度大于 5 的微博比率不到 2%，呈现急剧下降趋势。在含有热门主题微博中，下降趋势较为平缓，转发路径较长的微博比率较大，转发路径大于 5 的微博比率达到 12%以上。

图 3-14 给出了不同类型主题微博的总峰值持续时间分布，所有微博的总峰值时间中，前 4 个所占比率最大且不含热门主题的微博比率最大，占到总数的 98%。含有热门主题的微博下降趋势较为平缓，并且还存在小部分总峰值时间较长的微博，最大达到 15 个单位时间，这说明了此类型热门主题处于高度关注的时间较长。

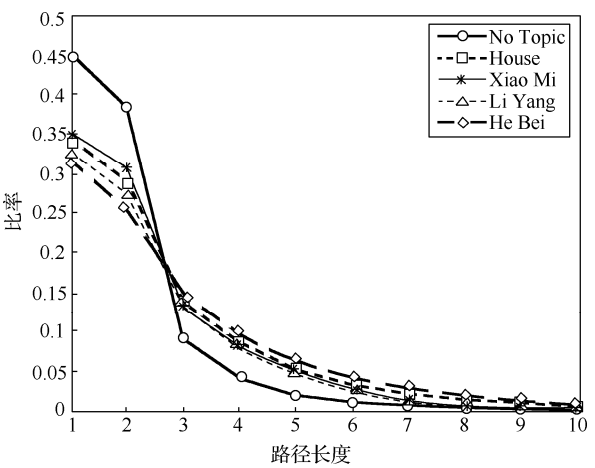


图 3-13　不同类型主题微博的转发路径分布

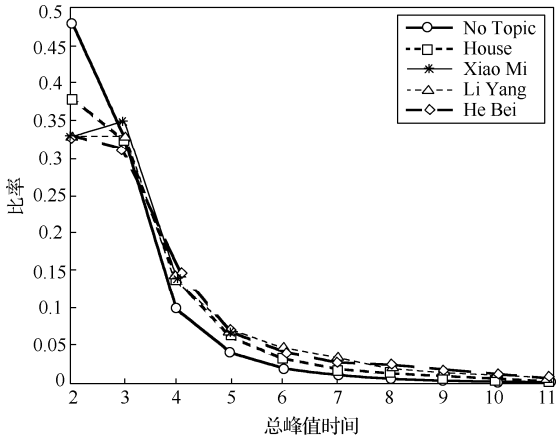


图 3-14　不同类型主题微博的总峰值时间分布

表 3-13 给出了不同类型主题微博的存活时间，所有微博的存活时间都呈现下降趋势，但不同类型主题的微博下降幅度有所不同。在不含热门主题的微博中，存活时间较短，存活时间在 3 天内的微博比率达到 75%，3 天后迅速下降，之后下降幅度又变缓，存在较长的拖尾，这说明存在一部分存活时间较长的不含热门主题微博。在含有热门主题的微博中，存活时间较长，尤其是河北大学飙车致死事件，存活时间低于 3 天的微博比率只有 60%左右，与不含热门主题的微博相差 15%左右，而存活时间超过 100 天的微博比率达到 9.45%，几乎是不含热门主题的微博的两倍，这说明此类主题的微博存活时间长，具有很强的生命力。另外，房价问题、李阳家暴事件等热门主题的微博同样具有较强的生命力，而小米手机发布事件的微博存活时间与不含热门主题的微博相似，说明此类微博的生命力一般。

表 3-13　不同类型主题微博的存活时间

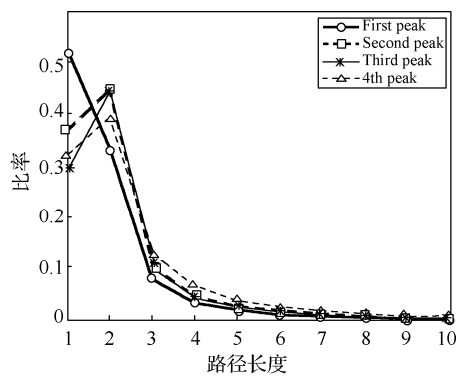
	不含主题		房价问题		小米手机发布		李阳家暴事件		河北大学飙车致死事件	
天　数	数目	比率	数目	比率	数目	比率	数目	比率	数目	比率
1	11 437	47.55	2 369	38.31	2 109	44.83	1 110	40.20	524	34.13
2	5 477	22.77	1 362	22.02	1 067	22.68	628	22.74	329	21.43
3	1 126	4.68	372	6.01	252	5.35	165	5.98	87	5.67
4～10	1 995	8.29	684	11.06	451	9.59	275	9.96	180	11.73
10～100	2 873	11.95	937	15.15	586	12.46	398	14.42	270	17.59
>100	1 145	4.76	461	7.45	239	5.09	185	6.70	145	9.45
总共	24 051	100	6 185	100	4 704	100	2 761	100	1 535	100

表 3-14 给出了不同峰值转发路径长度分布，它有两个显著的特征：一是在相同的主题微博中处于后面峰值的微博比处于前面峰值的转发路径长度要长，这说明随着时间的增加，微博转发路径在增长，深度在增加；二是在不同峰值时段含有热门主题微博的转发路径长度比不含热门主题的微博要长，说明含有热门主题的微博具有更大的影响力。

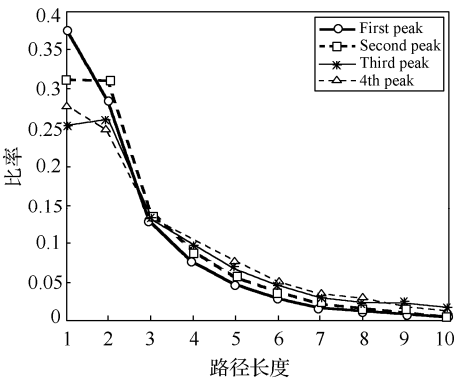
表 3-14　不同峰值转发路径长度分布

主 题 类 型	第一个峰值	第二个峰值	第三个峰值	第四个峰值
不含任何主题	1.78	2.01	2.32	2.39
房价问题	2.51	2.71	3.53	3.30
小米手机发布	2.46	2.65	3.71	3.45
李阳家暴事件	2.69	2.92	4.09	3.80
河北大学飙车致死事件	2.76	3.04	4.04	3.64

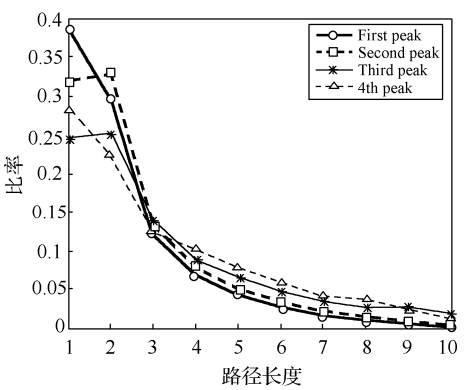
图 3-15 给出了 5 种不同类型微博在不同峰值时段转发路径长度比率分布。在不含热门主题的微博中，第一个峰值转发路径长度为 1 的微博比率最大，达到了50%，其他峰值转发路径长度为 1 的微博比率下降幅度较大，这说明在不含有热门主题的微博中，第一个峰值的转发方式与其他峰值不同。例如，在第一个峰值中大多数用户可能直接转发原始微博，而在其他峰值中用户转发的微博可能来自于其他用户。在含有热门主题的微博中，各个峰值的转发路径比率曲线变化比较相似，这说明了不同峰值时段的转发方式比较相似。



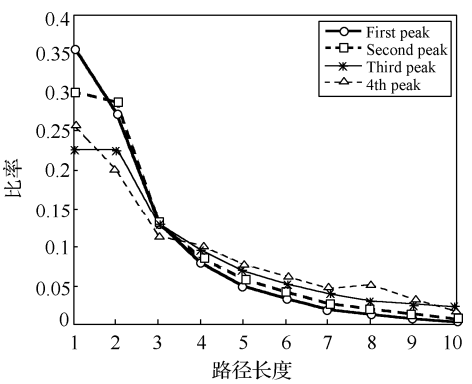
（a）不含热门主题微博



（b）房价问题事件

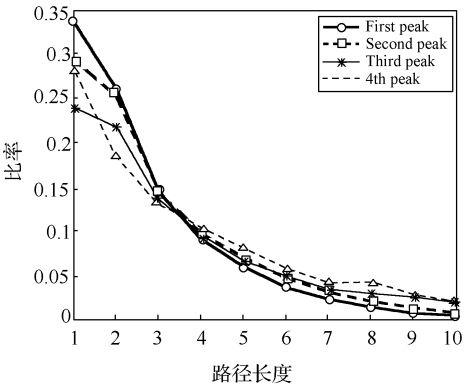


（c）李阳家暴事件



（d）小米手机发布事件

图 3-15　不同类型主题微博在不同峰值转发路径比率分布



（e）河北大学飙车致死事件

图 3-15　不同类型主题微博在不同峰值转发路径比率分布（续）

通过以上分析可知，不同类型的主题对微博转发时间序列多峰值有较大影响，具体表现在热门主题微博的转发路径长度、总峰值时间、存活时间以及峰值转发路径长度比不含热门主题的微博更长，这说明热门主题的微博影响力大、关注时间长以及生命周期长，并且不同热门主题的微博表现也不完全相同。

2）微博用户类型分析

用户类型包括超级用户、次超级用户和一般用户。不同类型用户的微博特征如表 3-15 所示，超级用户的微博转发路径长度、总峰值时间以及存活时间都较短，与一般用户微博相比，差距比较明显。超级用户的微博转发路径长度只有 1.86，大约是一般用户微博转发路径长度的 61%，而超级用户的存活时间大约是一般用户存活时间的37%。次超级用户则处于超级用户与一般用户之间。

表 3-15　不同类型用户的微博特征比较

用户类型	Average(P_l)	Average(T_p)	Average(T_d)(days)
超级用户	1.86	2.73	10.63
次超级用户	2.17	3.05	18.34
一般用户	3.04	3.77	28.46

图 3-16 给出了不同类型用户微博转发路径长度分布，随着转发路径长度的增加，所有微博比率都呈现下降趋势，不同类型用户的微博比率下降幅度有所不同，在超级用户和次超级用户中，路径长度短的微博比率较大，路径长度小于 2 的微博比率接近 80%，并且随着路径长度增长微博比率下降明显。在一般用户中，路径长度短的微博比率下降明显，路径长度小于 2 的微博比率下降幅度达到 25%，而路径长度长的微博比率有所上升，整个曲线较为平缓。

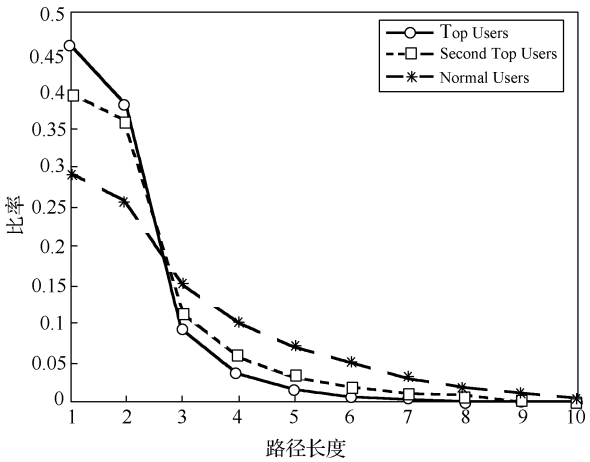


图 3-16　不同类型用户微博转发路径长度分布

图 3-17 给出了不同类型用户的微博总峰值时间分布，随着峰值持续时间增长，所有微博比率迅速下降，且总峰值时间中前 4 个所占比例最大。与图 3-16 相类似，超级用户的下降幅度比较剧烈，而一般用户的下降幅度较为平缓。

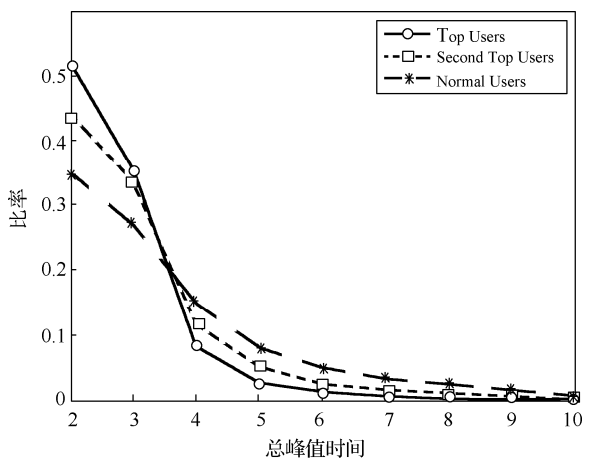


图 3-17　不同类型用户的微博总峰值时间分布

表 3-16 给出了不同类型用户微博的存活时间分布及比率，随着存活时间的增长，所有微博比率都呈现下降趋势，但不同类型用户的微博比率下降幅度有所不同。在超级用户中，微博的存活时间都较短，存活不超过 1 天的微博比率达到 60.25%，大大超过了平均值；存活超过 3 天的微博比率较少，不足 16%。在一般用户中，情况则不同，存活时间短的微博比率大幅度减少，而存活时间较长的微博比率较高，例如存活不超过 1 天的微博比率下降到

23.21%，存活超过 3 天的微博比率达到 54%，存活 100 天以上的微博比率高达 8.74%。由此可见，超级用户的微博存活时间分布与一般用户完全不同，次超级用户则处于超级用户与一般用户之间。

表 3-16　不同类型用户微博的存活时间分布及比率

	超级用户		次超级用户		一般用户	
天　数	数　目	比　率	数　目	比　率	数　目	比　率
1	10 055	60.25	4 964	43.19	2 334	23.21
2	3 998	23.96	2 633	22.91	1 930	19.20
3	461	2.76	608	5.29	840	8.36
4～10	679	4.07	1 040	9.05	1 717	17.08
10～100	944	5.66	1 602	13.94	2 352	23.40
>100	551	3.30	646	5.62	879	8.74
总数	16 688	100	11 493	100	10 052	1000

表 3-17 给出了用户在不同峰值时段的转发路径长度，它有两个特征：一是在相同类型用户中微博后一个峰值的转发路径长度比前一个峰值要长，尤其在超级用户中，这一特征更加明显，例如第四个峰值时的转发路径长度比第一个峰值时的转发路径长近 50%，这说明超级用户在不同峰值的转发方式是不同的；二是在不同峰值时段一般用户的微博转发路径长度都比超级用户要长，说明一般用户的微博传播深度更深一些。

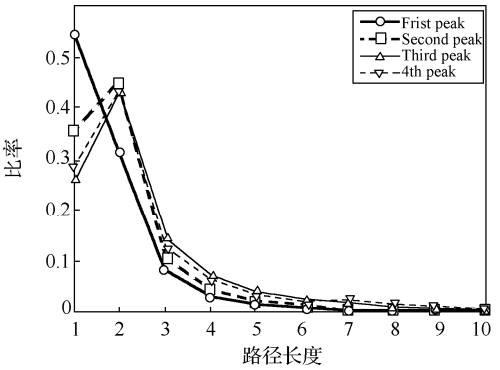
表 3-17　用户在不同峰值时段的转发路径长度

	第一个峰值	第二个峰值	第三个峰值	第四个峰值
超级用户	1.71	2.01	2.48	2.53
次超级用户	2.02	2.23	2.95	2.86
一般用户	2.99	3.03	3.22	3.26

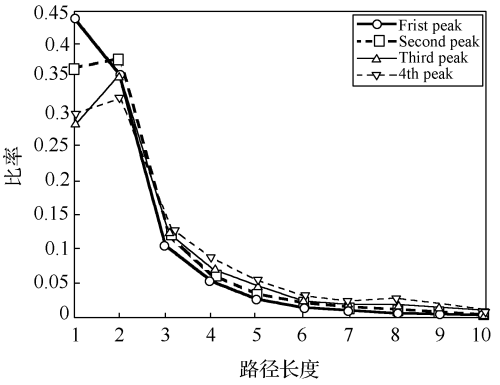
图 3-18 中给出了三种类型用户在不同峰值时段转发路径长度比率分布。在超级用户中，第一个峰值转发路径长度为 1 的微博比率很大，达到 50%，其他峰值转发路径长度为 1 的微博比率下降幅度较大，这再次说明了超级用户在第一个峰值的转发方式与其他峰值的转发方式不同。在一般用户中，不同峰值的转发路径比率变化比较相似，几乎重叠，这说明了用户的微博在不同峰值的转发方式基本相似。由此可见，不同类型用户的微博产生多峰值的原因是不同的。

以上数据分析表明，在不同类型的用户中，多个峰值的微博具有不同的特征，超级用户的微博转发数量较多，在短时间内快速地转发，其转发路径长度、总峰值时间以及存活时间都较短，因此超级用户的微博生命力偏短，可以看作一种“爆炸式”传播方式。在一般用户中，虽然微博转发数量较少，转发速度较慢，但是转发路径长度、总峰值时间以及存活时间都较长，对路径长度大于 2 的用户影响力更大，因此可以看作一种“蔓延式”的传播方

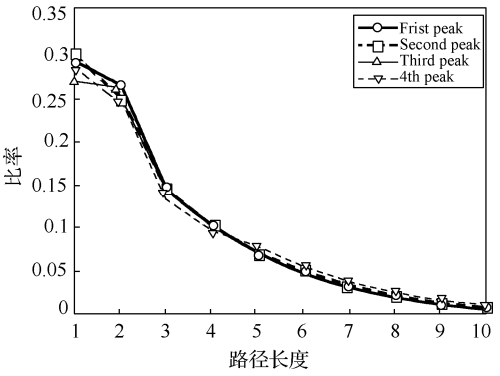
式。而次超级用户兼有两者的特征，既具有超级用户的“爆炸式”传播方式，也具有一般用户的“蔓延式”传播方式。



（a）超级用户



（b）次超级用户



（c）一般用户

图 3-18　不同类型用户在不同峰值时段转发路径长度比率分布

对于超级用户和一般用户的不同微博转发方式，主要是由微博网络的特殊性所决定的，他们在网络中所担当的角色不同。超级用户类似于媒体网络中的媒体，向用户传播信息主要依靠基于服从权威所形成的影响力，传播速度很快，但是持续时间较短。一般用户类似于社交圈子的朋友，用户之间的信息传播更多依靠亲戚朋友的口口相传的信任关系，这种传播方式速度较慢，但持续时间较长，在朋友圈子中更有影响力。

3.5　微博意见领袖识别

意见领袖又称舆论领袖，是指在信息传播网络中经常发表意见并具有相当影响力的“活跃分子”，他们在信息制造和传播过程中发挥着重要的作用，由他们将信息扩散给受众，在意见领袖的引导和影响下，局部意见可能演化为网络舆论。

统计数据显示，很多网民并不直接发表意见，而是通过关注和转发意见领袖的信息来表达自己的态度和倾向性，即所谓服从权威现象。通过意见领袖发表引导性信息来影响网民，可以有效地触发整个网络舆论的影响力。因此，意见领袖在推动信息传播、引导网络舆论中发挥着重要的作用。

在新浪、腾讯、网易等微博平台上，一些经过实名认证的高级账户，即贵宾账户（VIP）拥有众多粉丝，粉丝数量通常达到几十万，这样的微博用户被网民称为“网络大 V”。网络大 V 相当于意见领袖，具有很大的影响力，他们的一次转发就会使得一条微博迅速火起来，成为网络热点话题，引导着网络舆论走向。要分析网络大 V 或意见领袖在网络舆情中所起的作用，首先需要解决微博意见领袖识别问题。

下面介绍一种基于节点权重的微博意见领袖识别方法。

3.5.1　识别方法

基于节点权重的微博意见领袖识别方法的基本思路是根据网络拓扑特性，将网络抽象成一种有向图，通过分析节点之间的结构关系，计算每个节点的权值，节点权值越大，成为意见领袖的可能性就越大。因此，可以将意见领袖识别问题归结为如何计算节点权重问题。

首先将微博网络抽象成一种有向网络图 $G=(E, V)$，每个用户构成网络中的节点，用 E 表示节点关系集合；用户之间的关系构成节点之间的边，用 V 表示节点集合。由于每个用户拥有的朋友和粉丝数量不同，因此各个节点具有不同的权值，节点权值越大，说明该节点的影响力越大，成为意见领袖的可能性也就越大。在计算节点权重时，需要考虑到节点拥有的粉丝数量、节点连接关系以及交互关系等多种因素，以提高计算效率和精确度。

一个有效粉丝集合 Ef(u)定义如下：

$$\mathrm{Ef}(u)=\{v \mid v\in \mathrm{Follower}(u)\cap \mathrm{Response}(u)>\delta\} \tag{3-12}$$

式中，δ 是非负常数阈值，表示节点 u 的粉丝节点 v 对节点 u 反馈的门限，超过该阈值且属

于节点 u 的粉丝的节点才能算作有效粉丝。

由连接关系所产生的节点权值 IRL(u_i)的计算方法如下：

$$\mathrm{IRL}(u_i)=\frac{\sigma}{N}+(1-\sigma)\sum_{u_j\in \mathrm{Follower}(u_i)}\frac{\mathrm{IRL}(u_j)}{L(u_j)} \tag{3-13}$$

式中，IRL(u_i)表示节点 u_i 连接关系产生的节点权值，Follower(u_i)为节点 u_i 所有粉丝集合，$L(u_j)$为节点 u_j 粉丝数目，σ 是介于 0 和 1 的阻尼系数，N 为网络图中的总节点数。

由节点交互关系所产生的节点权值 IRTR(u_i)的计算方法如下：

$$\mathrm{IRTR}(u_i)=\sum_{t_j\in \mathrm{Tweet}(u_i)}\frac{\sum_{u_j\in \mathrm{Response}(t_j)}|N_s(u_j)-N_\mu(u_j)|}{|A|} \tag{3-14}$$

式中，IRTR(u_i)表示节点 u_i 的节点权值，Tweet(u_i)为用户 u_i 帖子集合，$|A|$表示所有具有交互情况的帖子集合，$N_s(u_j)$是节点 u_j 针对帖子 t_j 的响应次数，$N_\mu(u_j)$为响应平均值，Response 包括用户转帖、回帖、评论和收藏。

节点综合权值 $IR(u_i)$的计算方法如下：

$$IR(u_i)=(1-\beta)\times(IRL(u_i)+\beta)\times IRTR(u_i) \tag{3-15}$$

式中，参数 β（$\beta\in[0, 1]$）主要决定连接关系和节点交互关系两个因子在节点权值计算中所处的地位。当 β 较小时，节点权值主要由连接关系决定，特别当 $\beta=0$ 时，则完全由连接关系计算权值。

综上所述，该方法的具体算法如下。

算法 3-3　基于多连接的节点权重算法 （Multi-Link）

（1）利用网络爬虫工具，从互联网中采集实际的微博网络数据，提取其中的节点、连接等网络拓扑信息存入数据库待处理；

（2）构建有向网络图 $G=(V, E)$；

（3）利用式（3-12）计算有效粉丝集合 Ef(u)；

（4）利用式（3-13）计算由连接关系所产生的节点权值 IRL(u_i)；

（5）利用式（3-14）计算由节点交互关系所产生的节点权值 IRTR(u_i)；

（6）利用式（3-15）计算节点综合权值 IR(u_i)；

（7）计算网络图中所有节点的综合权值，并按综合权值由大到小排序，选取综合权值较大的 n 个节点，作为意见领袖的候选对象。

由于该方法在计算节点权重时考虑了节点粉丝数量、节点连接关系以及交互关系等多种因素，因此称为基于多连接（Multi-Link）的节点权重计算方法，它从计算效率和准确度两个方面改进了现有方法的不足，一方面，通过定义有效粉丝集合，将没有或拥有少量粉丝的节点排除掉，他们成为意见领袖的可能性极小，因为意见领袖或高权值节点必然拥有大量

粉丝，这样就可大幅度减小网络图规模，有利于提高计算效率。另一方面，在计算节点权值时，不仅考虑了由粉丝产生的连接关系，还考虑了帖子的发布、转发、回复以及收藏等所产生的节点交互关系，因此提高了计算精确度。

3.5.2　算法验证

由于意见领袖的识别被量化成网络中节点权值序列，在这个序列中排名靠前的节点可认为是网络中的意见领袖。目前还没有用于衡量意见领袖识别效果的标准，学术界主要采用算法对比方式来评价意见领袖识别效果。

下面通过实验数据对基于多连接（Multi-Link）算法和基于网络拓扑特性（Topological-based）算法的性能进行三种统计学方法的测试和对比，三种统计学方法包括 T-Test 检验、Kendall tau Rank 检验和 Spearman Rank 检验。

1．实验数据集

实验数据集是从互联网中采集的真实社交网络数据，其数据集来源及规模如表 3-18 所示。

表 3-18　数据集来源及规模

数据集来源	帖子数目	用户节点
优酷	330 232	5 655
Youtube	4 945 382	1 138 499
其他论坛	53 332 256	1 000 000
新浪微博	2 370 238	350 747

2．T-Test 检验

T-Test 检验也称 Student-t 检验，主要用于检验样本空间较小（例如 n<30）、总体标准差 σ 未知的正态分布数据。

首先使用 Multi-Link 算法和 Topological-based 算法分别对 10 万个新浪微博用户节点进行意见领袖识别，得到前 100 位节点权值排名靠前的用户节点，然后对这 100 个用户节点使用 T-Test 检验，得到这些节点的 P-Value 分布，图 3-19 和图 3-20 分别给出了 Multi-Link 算法和 Topological-based 算法的 T-Test 检验的 P-Value 分布。

图中直线标识了 P-Value = 0.05 即 5%的分割线，可以看出，节点的 P-Value 值主要集中在该直线以下，即通过 T-Test 检验发现，两种算法计算的节点领袖权值具有较高可信度，能够代表网络中的意见领袖。

3．Kendall-tau Rank 检验

在统计学中，肯德尔相关系数（Kendall-tau）是用来测量两个随机变量相关性的统计值，用 τ 表示其值。一个肯德尔检验是一个无参数假设检验，它使用计算得到的相关系数去检验两个随机变量的统计依赖性。τ 的取值范围为–1～1，当 τ 为 1 时，表示两个随机变量

拥有一致的等级相关性；当 τ 为–1 时，表示两个随机变量拥有完全相反的等级相关性；当 τ 为 0 时，表示两个随机变量是相互独立的。τ 的计算公式如下：

$$\tau = \frac{n_e - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{3-16}$$

（1）如果排列双方的排名是完美的（即两个排名是相同的），τ 值为 1；

（2）如果两个排列之间的分歧排名是完美的（即一个排名为扭转其他），τ 值为–1；

（3）对于所有其他 τ 值在–1 和 1 之间的排列，增加值意味着增加排列之间的排名。

根据计算结果，Multi-Link 算法和基于拓扑的算法之间的 τ 值为 0.9107，说明这两种算法具有很高的一致性。

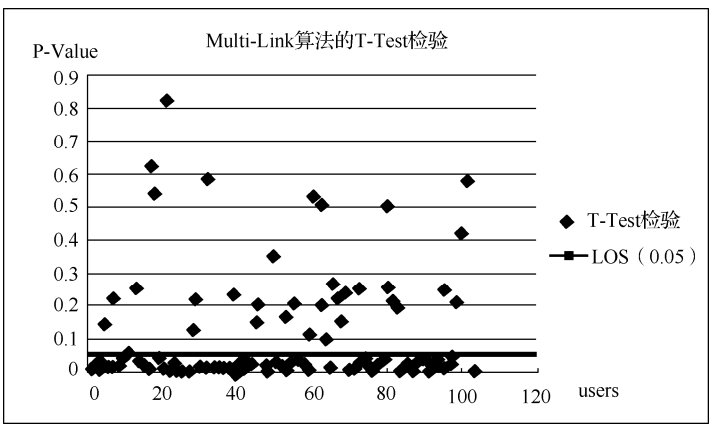


图 3-19　Multi-Link 算法的 T-Test 检验的 P-Value 分布

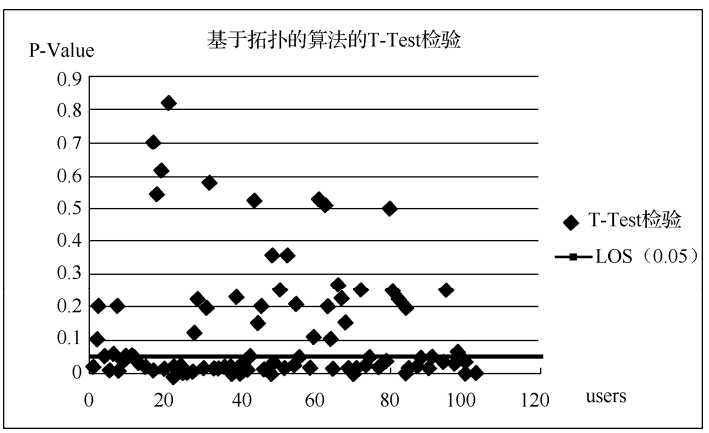


图 3-20　基于拓扑的算法的 T-Test 检验的 P-Value 分布

4．Spearman Rank 检验

在统计学中，斯皮尔曼等级相关系数（Spearman Rank）用来估计两个变量 X、Y 之间的相关性，其中变量间的相关性可以使用单调函数来描述，并用 ρ 表示其值。如果两个变量取值的两个集合中均不存在相同的两个元素，那么，当其中一个变量可以表示为另一个变量的很好的单调函数（即两个变量的变化趋势相同）时，两个变量之间的 ρ 值范围在–1 到 1 之间。

假设两个随机变量分别为 X、Y（也可以看作两个集合），它们的元素个数均为 N，两个随机变量取的第 i（$1 \leqslant i \leqslant N$）个值分别用 X_i、Y_i 表示。对 X、Y 进行排序（同时为升序或降序），得到两个元素排行集合 x、y，其中元素 x_i、y_i 分别为 X_i 在 X 中的排行以及 Y_i 在 Y 中的排行。将集合 x、y 中的元素对应相减得到一个排行差分集合 d，其中 $d_i = x_i - y_i$，$1 \leqslant i \leqslant N$。随机变量 X、Y 之间的 ρ 值可以由 x、y 或者 d 计算得到，其计算方式如下：

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3-17}$$

表 3-19 给出了 7 种算法之间的 Spareman Rank 值，从表 3-19 可以看出，Multi-Link 算法和基于拓扑的算法具有较高的 Spareman Rank 值，序列一致性较高，说明 Multi-Link 算法和基于拓扑的算法在意见领袖识别上表现出较好的能力。

表 3-19　各个算法的 Spareman Rank 值

	A	B	C	D	E	F	G
A	-	0.947	0.953	0.821	0.773	0.873	0.921
B	0.947	-	0.924	0.726	0.842	0.630	0.876
C	0.953	0.924	-	0.843	0.782	0.760	0.844
D	0.821	0.726	0.849	-	0.854	0.861	0.910
E	0.773	0.842	0.782	0.854	-	0.910	0.842
F	0.873	0.623	0.760	0.861	0.910	-	0.891
G	0.921	0.876	0.844	0.910	0.846	0.891	-

注：各个字母所代表的算法，A：Topological；B：Topic；C：Multi-Link；D：PageRank；E：HITS；F：TwitterRank；G：InfluenceRank。

5．算法性能对比

使用准确率和召回率来评价意见领袖识别算法性能，准确率（P）和召回率（R）的计算公式如下：

$$P = \frac{A}{A+B}, \quad R = \frac{A}{A+C} \tag{3-18}$$

式中，A 为识别出的意见领袖数目，B 为识别出的非意见领袖数目，C 为未识别出的意见领袖数目。

由于在意见领袖识别中还没有标准来衡量是否识别出全部的意见领袖，因此在计算准确率和召回率时通常以经验的意见领袖数目来近似真实的意见领袖数目。表 3-20 为各种算法的准确率、召回率及平均节点处理时间。

表 3-20　各种算法的准确率、召回率及平均节点处理时间对比

算　法	准　确　率	召　回　率	时间（min）/10 万节点
出度	0.622	0.571	0.143
出度/入度结合	0.673	0.654	0.236
ThreadRank	0.861	0.823	3.376
InfluenceRank	0.847	0.817	2.813
TwitterRank	0.905	0.885	2.763
Topological-based	0.924	0.912	2.215
Topic-based	0.925	0.907	2.860
Multi-Link	0.918	0.894	1.316

注：时间测试是在包含 10 万个用户节点的真实数据环境下得到的结果。

由于 Multi-Link 算法采用微博网络拓扑结构中连接关系与节点交互相结合的计算方法，降低了网络节点规模，从而提高了计算速度，同时准确率和召回率也有显著的提高。图 3-21 和图 3-22 分别给出了不同算法的准确率和召回率以及计算时间比较。

从图 3-21 可以看出，在测试数据集上，Multi-Link、Topological-based 及 Topic-based 等算法的召回率和准确率比较高，与 TwitterRank 算法基本相当，比常见的出度和出度/入度结合算法更好，而出度和出度/入度结合算法的召回率和准确率都比较低。

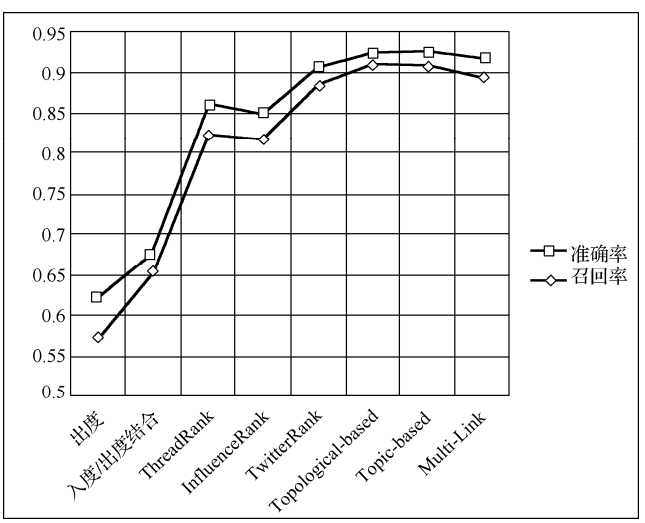


图 3-21　不同算法的准确率和召回率

从图 3-22 可以看出，出度和出度/入度结合两种算法的计算时间比较短，因为在计算过程中，这两种算法没有考虑其他的附加条件，算法比较简单，但召回率和准确率都比较低。而其他算法由于考虑了更多的修正因素，因此计算时间稍长。相比之下，Multi-Link 算法的计算时间处于中等水平。

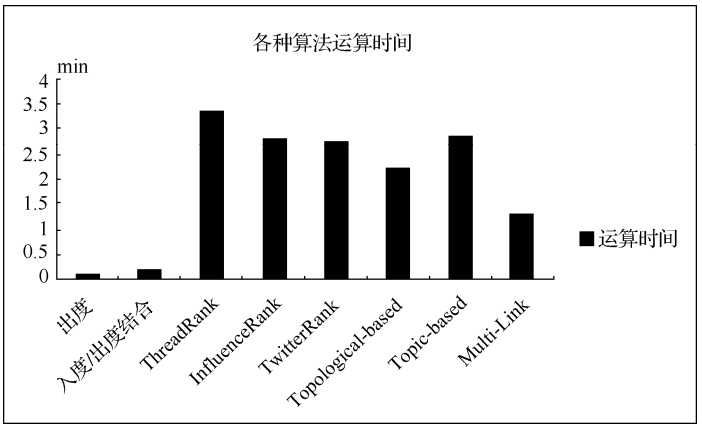


图 3-22　不同算法的计算时间

根据 T-Test、Kendall-tau 和 Spareman Rank 三种统计学检验方法的对比实验结果，表明 Multi-Link 算法具有较高的意见领袖识别能力，与 Topological-based、Topic-based 等算法具有一致性。

根据算法的准确率、召回率以及计算时间的实验结果，表明 Multi-Link 算法不仅在准确率和召回率上表现良好，并且比 Topological-based、Topic-based 等算法的计算时间要短，这对于处理海量网络数据来说是至关重要的。

因此，从意见领袖识别能力、准确率和召回率、计算时间等综合指标来看，Multi-Link 算法更具优势。

第4章

网络论坛舆情传播机制

4.1 引言

网络论坛是一种为用户提供信息交流平台的网络应用系统，网络论坛也称为电子公告板BBS（Bulletin Board System），最早是用来发布股市价格等信息的，当时的 BBS 功能比较简单，连文件传输功能都没有。随着计算机技术和网络技术的发展，以及网络信息交流需求的驱动，BBS 不断发展壮大，现在的网络论坛几乎涵盖了社会生活的方方面面，每个用户都可以找到自己感兴趣或者需要了解的专题性论坛，包括综合性门户网站和功能性专题网站等各类网站也都开设了自己的论坛，以促进网民之间的交流，增强网民的互动性。

网络论坛属于传统的网络信息交流平台，随着社交网络、微博网络等新型网络信息交流平台的广泛应用，网络论坛的用户数量有所下降，尽管其网民数和使用率不如微博、社交网络高，但网络论坛所具有的多元化、开放性、匿名性及互动性，仍然是广大网民发表言论、获取信息的重要网络平台，用户数量还是比较庞大的。

网络论坛以多元化、开放性、匿名性及互动性为特色，为网民提供了发表言论、获取信息的网络信息交流平台。在网络论坛中，网民就某个主题通过发帖、观看和回帖进行信息交流和互动，在信息交流过程中，某些话题的帖子受到网民的高度关注，点击量和回帖数非常大, 形成较大的影响力，这种帖子称为热帖，热帖在观点传播和舆论形成过程中起到重要的推动作用。可见，网民通过发帖和回帖发表意见，参与观点传播和舆论形成，对于推进社会进步和政治民主起到了积极的作用，成为网络舆情的主要来源。同时，在网络舆情中存在着通过网络炒家人为炒作出来的虚假网络舆情，容易产生错误的舆论导向，危及政府的公信力，引发社会群体性事件等问题。

网络论坛舆情问题引起了社会和学术界的极大关注，研究人员通过建立相应的数学模型，对网络论坛的信息传播特性、网络舆情检测、意见领袖发现、网络水军识别等问题进行了研究，找出其中的内在规律，为快速检测网络舆情、识别虚假网络舆情、抑制网络谣言传播提供科学依据和解决方案。

本章主要介绍网络论坛的舆情形成模型、意见领袖识别、网络水军热帖检测、网络水军账号识别等内容。

4.2　网络论坛舆情形成模型

4.2.1　网络论坛结构

通常，一个网络论坛由很多板块（Board）构成，组织形式为主题（Thread）。论坛中某一板块用户间的讨论首先是以第一个作者发出一个主帖（Entry Post）开始的，该主帖包含一个唯一的标题（Title）和入口（Entry），然后由其他用户围绕该主帖通过发一个或多个包含相应内容的回帖（Post）来进行的。论坛的话题（Topic）是一个或多个讨论类似内容的网络论坛的主题（Thread）的集合。图 4-1 为网络论坛结构图。

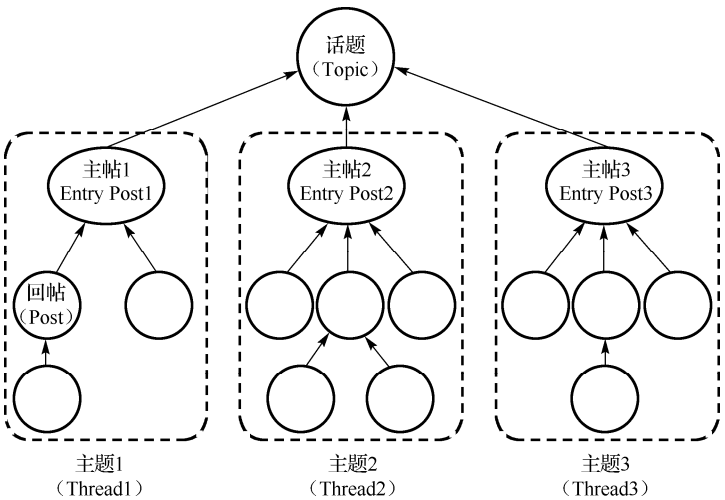


图 4-1　网络论坛结构

（1）主帖：主帖是作者就某个主题进行的详细论述，它包括唯一的标题和入口、作者 ID、发表时间、详细的主题论述内容，在每一个主题中，主帖是唯一的。

（2）标题：主帖的标题，也是所在主帖的标题，由发表主帖的作者首先提出，代表着所在主题将要描述的事件对象的主要内容，主帖所有的回帖基本都是围绕该标题展开的。主帖标题基本能够概括事件的主题。

（3）回帖：论坛用户对于某一主题下的相关帖子（包括主帖或主帖的回帖）的回复，它包括作者 ID、回复时间、回复内容。在每一个主题中，回帖可能不止一个。

（4）主题：论坛的信息组织形式，一个主题由唯一的主帖、唯一的主帖标题、唯一的

主帖入口和众多回帖组成。同一话题可能包含多个从该话题不同角度进行讨论的主题。

（5）话题：话题是由一个或者多个讨论相似内容的主题组成的集合，包含了围绕现实社会某一事件对象的全部属性。同一话题的不同主题在论坛中持续的时间也不尽相同，它们共同组成了该话题完整的生命周期。

4.2.2 舆情形成模型

对于网络论坛观点传播和舆情形成模型，比较典型的有 Sznajd 模型和 French-DeGroot 模型等，后来的很多模型都是这两种模型的扩展。Sznajd 模型是一个基于一维空间的舆情形成模型，认为节点是排列在一维空间上的点，并假定节点可以选择其中小数量的离散观点，由于该模型对选择过程给出了较好的解释而受到了广泛关注，并被推广到小世界网络和无标度网络。French-DeGroot 模型认为节点的观点可以在一个任意维度和结构的空间上延伸，观点在吸引力驱使下实时演化，即节点在观点空间中变换立场，趋向于其他节点观点比较集中的领域，该模型能够比较真实地反映现实网络中节点观点的倾向问题。

Sznajd 模型和 French-DeGroot 模型虽然在一定程度上表达了观点传播和舆情形成的主要特征，但也存在以下问题：一是 Sznajd 模型认为人们表达的观点只有两种：支持和反对，不能真实反映节点观点的模糊性和连续性；二是 French-DeGroot 模型假定网络中边的权值是不随时间变化的固定值，不能真实反映节点间联系的亲密程度和时变特性。

下面给出一个基于节点影响力的网络舆情形成模型，在 French-DeGroot 模型的基础上，考虑到网络中节点观点变化的连续性以及节点间不同的连接强度及其时变特性，能够比较真实地反映网络论坛中的舆情形成过程。

在网络论坛中，用户之间通过发帖和回帖进行信息交互。为了描述网络论坛中的观点传播和舆情形成过程，以网络论坛中的用户为节点、节点间相互连接为边来构建一个无向权值网络图，称为论坛网络，即：

$$G=(V,E,W) \tag{4-1}$$

式中，V 为论坛网络的节点数；E 为连接所有节点的边的集合，表示节点观点可能的传播路径；$W=[w_{ij}]$为边的权值矩阵，定义为节点 i 对与其有连接的各个节点的影响力而构成的影响力矩阵。

这里假设论坛网络是一个封闭的网络，舆情在论坛网络中产生，并仅在论坛网络中传播。同时假设论坛网络中的节点数是不变的，每一个节点在 t 时刻都有一个用实数表示的观点值 $x(t)$。

在论坛网络中，节点间的连接因它们观点的不同具有不同的强度和内涵，并且是随时间动态变化的。为了描述这种特征，假设 t 时刻网络中节点 i 和 j 的观点分别为 $x_i(t)$，$x_j(t)$，则节点间的观点距离 d_{ij}^t 可表示为：

$$d_{ij}^{t}=\left|x_{t}(t)-x_{j}(t),\forall i,j\in N,i\neq j\right|$$

令

$$w_{ij}^{t}=\frac{1}{d_{ij}^{t}}=\frac{1}{\max(d,\left|x_{i}(t)-x_{j}(t)\right|)} \tag{4-2}$$

式中，d 是一个非常小的正数，以代替 $x_i(t)=x_j(t)$时分母为零的情况。式（4-2）表明，两个节点间的观点距离越接近，彼此间的影响力就越大，这是符合实际的，例如社会中两个人的关系越亲密，当其中一个人遇到问题时就愿意从亲密的朋友那里寻求建议，该朋友对他的影响力也就越大。

为了描述问题，定义影响力矩阵 T 为 $n\times n$ 的非负矩阵，则对所有的 $i,j\in N$，$T_{i,j}^{i}\in[0,1]$ 表示 t 时刻节点 i 对节点 j 的观点影响权重，同时 T 是一行随机矩阵，即 $\sum_{j=1}^{n}T_{ij}^{t}=1$。对 $t\geqslant 0$，在 t+1 时刻节点间的影响力可表示为：

$$T_{ij}^{t+1}=\frac{w_{ij}^{t}}{T_{ii}^{t}+\sum_{k\in N_{-i}}w_{ik}^{t}} \tag{4-3}$$

式中，T_{ii}^{t} 为 t 时刻节点 i 的自我影响力，N_{-i} 为除去节点 i 以外的节点集。

在 t+1 时刻节点 i 的自我影响力可表示为：

$$T_{ii}^{t+1}=1-\sum_{k\in N_{-i}}T_{ij}^{t+1} \tag{4-4}$$

因此，一个舆情系统定义如下：

$$S=\{n,t,x(t),T,G_{t}(V,E,W)\} \tag{4-5}$$

式中，n 表示系统中的节点个数，$t=\{0,1,2,\cdots\}$为系统中离散的时间点，$x(t)=[x_1(t),\ x_2(t),\cdots,x_n(t)]$ 表示 t 时刻系统的观点剖面，$T=[T_{ij}]$ 是影响力矩阵，$G_t\ (V,E,W)$为 t 时刻系统的网络拓扑图。

舆情形成模型描述了论坛网络中用户观点随时间变化的趋势和网络舆情形成过程。设节点 $i\in V$ 在 $t\in T$ 时刻的观点为 $x_i(t)\in x\ (t)$，则节点 i 在 t+1 时刻的观点为 $x_i(t$+1)，可表示为：

$$x_{i}(t+1)=\sum_{j=1}^{n}T_{ij}^{t+1}x_{j}(t) \tag{4-6}$$

式中，$x_i(t$+1)为节点 i 在 t+1 时刻的观点值，T_{ij}^{t+1} 为节点 j 在 t+1 时刻对节点 i 的影响力值。

舆情系统 S 在 t+1 时刻的观点剖面可表示为：

$$x(t+1)=Tx(t) \tag{4-7}$$

式中，$x\ (t)=[x_1\ (t)x_1\ (t)\ \cdots\ x_n\ (t)]^{*}$为 t 时刻的观点剖面，$T=[T_{ij}]$为 $n\times n$ 的影响力矩阵。

这样，在论坛网络中，每一节点随着时间的推移总是不断地改变着邻居节点的影响力并更新自身影响力。同时也会根据邻居节点当前的影响力和邻居节点上一时刻的观点来更新自身观点。节点观点演化的本质就是基于节点影响力的节点观点迭代过程。

如果舆情系统 $S=[S[1],\cdots,S[n]]$中的任意两节点 i 和 j 均满足$|x_i(t)-x_j(t)|<\varepsilon$（$\varepsilon$ 是一个很小的实数），则舆情系统 S 收敛。

可见，在论坛网络中，当 $t\to\infty$时，节点观点在外部环境的持续影响下不断发生改变。如果所有节点观点在某一时刻 t^*均收敛于一定值 $S(t^*)$，则称该网络中的节点观点达到渐进一致或完全一致，形成网络舆情。

4.2.3 模型验证

下面通过仿真实验方法对论坛网络的舆情形成模型进行测试和验证。

1．仿真工具与步骤

在仿真实验方法中，采用 UCINET 6.0 软件作为论坛网络生成与分析工具，采用 MATLAB 7.0 软件来仿真论坛网络的观点传播和舆情形成过程。

仿真步骤如下：

（1）初始化仿真起始时间 $t=0$，仿真时间 t_N。并通过 UCINET 6.0 随机生成包含 n 个节点的论坛网络，为各个节点随机分配一个初始观点值 $x_i(t)\in[1,5]$ 以及自我影响的初始影响力值 $T_{ii}^t\in[0,1]$。

（2）按照式（4-3）和式（4-4）分别计算每一节点在 $t=t+1$ 时刻对邻居节点及自身的影响力，构建影响力矩阵 $T_{ij}^{t+1}(d=10^{-2})$。并按照式（4-7）计算在 $t+1$ 时刻系统的观点剖面值。

（3）对任意节点 i 和 j，如果在 $t=t^*$时刻满足$\left|x_i(t^*)-x_j(t^*)\right|<\varepsilon\,(\varepsilon=10^{-3})$，则结束仿真；否则，重复步骤（2）和步骤（3）。

2．模型有效性验证

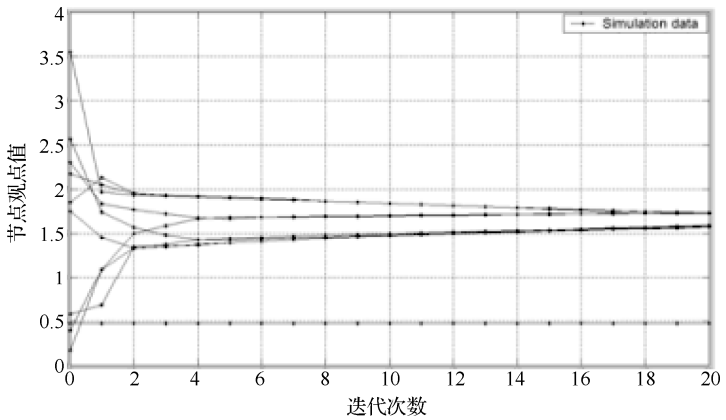
为了验证模型的有效性，选取节点数 n 分别为 10、30、50 和 70 的 4 个不同规模的论坛网络进行仿真，得到如图 4-2 所示的舆情形成过程图，图中 Y 轴代表各节点的观点值，X 轴代表迭代次数。从图 4-2 可以看出：

（1）无论网络规模如何变化，模型均能收敛。即论坛网络中存在直接或间接连接的节点通过相互作用，各自持有的观点逐渐发生改变，最终收敛于某一定值。在这个过程中，观点传播在起始几个周期变化梯度较大，随着时间推移变化趋缓，逐渐趋向一个稳定值，表明系统从非平衡态趋向平衡态。这与现实网络论坛中的观点传播趋势相一致，最初众人观点不一，随着彼此讨论、相互影响，最终观点趋于一致，形成舆情。

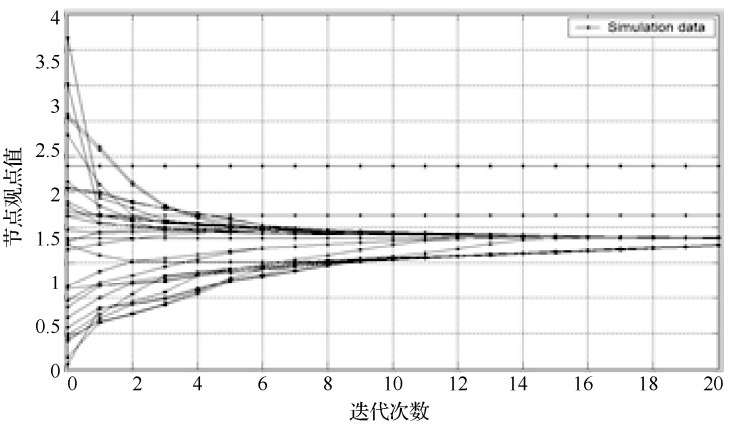
（2）观点最终形成两种不同的簇，即论坛网络中的节点最终形成两种持不同观点的群。这进一步验证了论坛网络具有的社区结构特性，即具有较高观点相似度值的节点可以划归于同一类型的社区，它们之间存在较多的连接；而社区与社区之间没有或有很少连接。

（3）对照网络图可以发现，通过较短的观点传播周期达到最终收敛值的节点存在更多的邻居节点。这与现实网络论坛中的情况相符，即邻居较多的用户因有更多的机会参与交互，其持有的观点更容易影响到其他用户，同时也更容易受到其他用户的影响，这些用户的观点状态趋于一致的速度更快。

（4）孤立节点因不参与交互而不会影响到其他节点和受其他节点的影响，其观点状态不会发生改变（在图 4-2 中其变化趋势为直线）。在现实网络论坛中，这类节点对应于只浏览帖子而不发帖的用户。

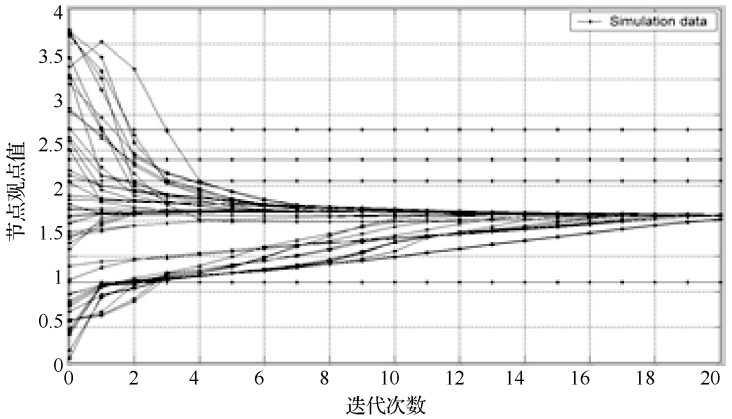


（a）节点 $n = 10$ 下的舆情形成过程

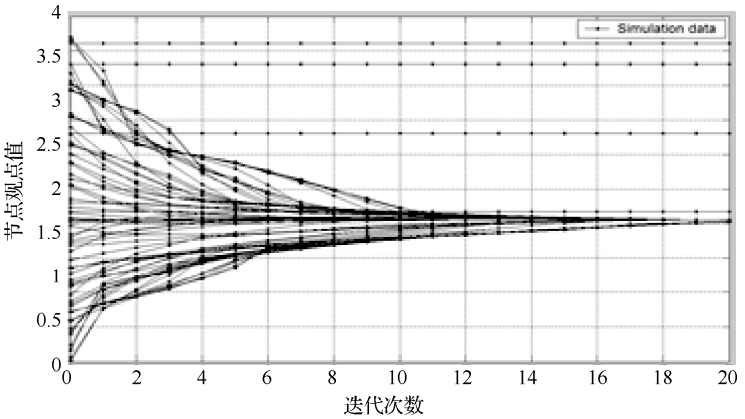


（b）节点 $n = 30$ 下的舆情形成过程

图 4-2　观点传播与舆情形成过程



（c）节点 $n=50$ 下的舆情形成过程



（d）节点 $n=70$ 下的舆情形成过程

图 4-2　观点传播与舆情形成过程（续）

（5）在节点数 $n=70$ 的论坛网络中，节点观点的收敛速度快于其他 3 个网络，说明网络规模并非是影响网络舆情形成的主要因素，网络规模大时，舆情形成速度可能更快。

4.3　网络论坛意见领袖识别

在网络舆论形成过程中，意见领袖发挥了积极的推动作用。在网络论坛中，大部分用户以浏览为主，对感兴趣的话题进行回复，他们的观点往往跟随意见领袖。在意见领袖的引导和影响下，局部观点或意见可能演化为网络舆情。因此，通过意见领袖来引导和控制网络舆情是十分重要的。要达到这一目的，首先需要解决网络论坛意见领袖识别问题。

在网络论坛意见领袖识别上，国内外研究者进行了广泛研究，提出了各种识别方法，包括简单统计测量法、影响力扩散模型、网页排名（PageRank）法、社会网络分析法等。

其中，社会网络分析法采用社会网络分析中的相关量化指标来发现论坛中有影响力的用户，具有准确、简洁和高效等优点。然而，由于现有的社会网络分析法主要关注于静态网络的结构和统计学特性，因而不能反映网络论坛中用户间交互关系的动态演变特性，如用户在论坛中的角色和权限会随时间的推移而发生变化等，对意见领袖识别的准确度产生一定的影响。

下面给出一种基于时间变化图的论坛意见领袖识别方法，将社会网络分析法和时间变化图相结合，提高了论坛意见领袖识别的准确度。

4.3.1　论坛有向网络图模型

将网络论坛的用户作为节点，一个用户对另一个用户的回帖看作施加一个影响，如 B 对 A 进行了回复，则 B 对 A 施加了一个影响。这样，根据论坛用户间的交互关系，可以将网络论坛抽象成一个论坛有向网络图，如图 4-3 所示。在一个时间周期内，用户间就某一主题发出一定数量的帖子（包括发帖和回帖），则可将图 4-3 转换成以帖子数为边权值的论坛有向权值网络图，如图 4-4 所示。

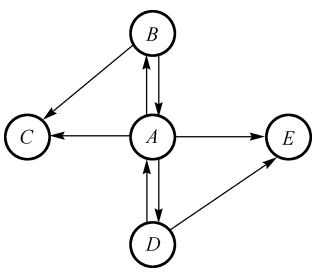


图 4-3　论坛有向网络图

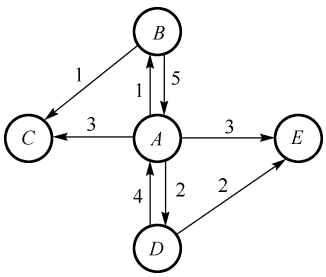


图 4-4　论坛有向权值网络图

事实上，基于论坛用户间交互关系的论坛网络是动态变化的，随着时间的推移，节点会不断地加入或离开网络，节点间的边会因此而发生变化，边的权值以及节点在网络中的角色和权限也会随之发生动态变化。图 4-5 为论坛网络动态演变过程，经过三个时间周期 T_1、T_2、T_3 的演化，形成一个有向权值网络图。

根据以上分析，基于论坛用户间交互关系的论坛网络图定义如下：

$$G=(V,E,W,T) \tag{4-8}$$

式中，V 为节点，表示网络论坛中的用户；E 为边，表示论坛用户间的交互关系；W为边的权值，表示用户间交互的帖子数量；T为网络图的生命周期，表示用户交互的持续时间。

生命周期 T可以分割成连续的子区间 $T=[t_0,t_1)$，$[t_1,t_2)$，…，$[t_k,t_{k+1})$，每一个区间$[t_k,t_{k+1})$用 T_k 表示，则在每一个 T_k 上的论坛网络图可以表示为 $G_k=(V[t_k,t_{k+1})$，$E[t_k,t_{k+1})$，$W[t_k,t_{k+1})$，$T_k)$，每一个时间窗口 T_k 对应的是论坛网络的一个即时快照。因此，一个论坛有向权值网络图可以表示为一连串在各个时间窗口内的图的序列，即 $SF(t)=G_0,G_1,\cdots,G_k$。

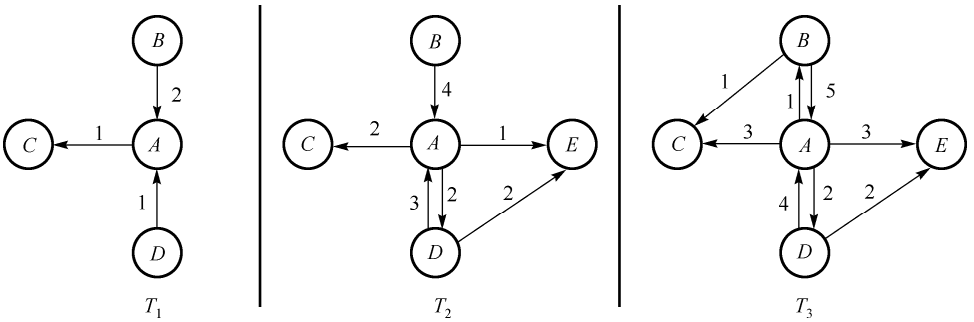


图 4-5　论坛网络动态演变过程

4.3.2　论坛意见领袖识别算法

意见领袖是在观点形成与传播中扮演重要角色的人，他们在一个网络中的特殊位置以及交流习惯来影响其他用户的观点，给那些搜寻信息的用户提供一种导向，具有较大的影响力。因此，意见领袖识别问题可以转化为如何计算用户影响力，根据用户影响力值识别出意见领袖。

研究发现，网络论坛中的意见领袖通常具有以下两方面的行为特征。

（1）意见领袖总是与论坛中多个用户存在直接联系；

（2）在一定时间周期内，意见领袖往往非常频繁地直接与论坛中多个用户发生交互，并频繁向多个用户发布和回复帖子。

依据网络论坛中意见领袖的行为特征，可以采用节点度和聚类指标来量化论坛用户的影响力。

在论坛网络中，使用节点度来衡量一个用户直接与其他用户交流的频繁程度，节点度可以进一步分为出度和入度，分别表示一个用户在某一时间周期内发出和接收的帖子数量。如果一个节点具有一个高出度值，表示该用户在网络论坛中发出了比较多的帖子，因此有更多的机会去影响其他人；如果一个节点具有较高的入度值和非常低的出度值，表示该用户是一个很少参与交流的不活跃用户。出度和入度可以用如下公式表示：

$$D_O(i)=\sum_{j\in N}\vec{\boldsymbol{e}}(i,j)w_{i,j}$$

$$D_I(i)=\sum_{j\in N}\vec{\boldsymbol{e}}(j,i)w_{j,i} \tag{4-9}$$

式中，i，j 分别代表图中的两个节点，$\vec{\boldsymbol{e}}(i,j)\in E$ 代表从节点 i 到节点 j 的一个有向边，$w_{i,j}\in W$ 代表有向边的权值，N 代表节点 i 的邻接节点集。

聚类是衡量一个用户与一个高度互联的用户群的亲密程度。在论坛网络中，节点的聚

类又分为引入聚类（Incoming-clustering）和外出聚类（Outgoing-clustering），分别表示一个用户在某一时间周期内向用户群发出或接收的帖子数量。一个节点具有较高的外出聚类值，表示该用户发出的帖子可以快速地在用户群内传播，并可通过用户群传播到用户群以外的更大范围。因此，一个具有高外出聚类值的用户具有较大的机会成为意见领袖。同理，如果一个节点具有较高的引入聚类值，表示该节点与较多的用户群存在连接，信息来源较多，有更多的机会来接受他人的意见。

外出聚类和引入聚类值可以用如下公式表示：

$$C_O(i)=\frac{\sum_{j\in N}D_O(j)}{D_O(i)\times(D_O(i)-1)}$$

$$C_I(i)=\frac{\sum_{j\in N}D_I(j)}{D_I(i)\times(D_I(i)-1)} \tag{4-10}$$

式中，j 表示用户群中的节点，$\sum_{j\in N}D_O(j)$ 和 $\sum_{j\in N}D_I(j)$ 表示用户群中节点间实际存在的边。

通过以上分析可以得出：具有高出度值和高外出聚类值的用户具有很大的影响力，在某一时间窗口内成为意见领袖的可能性很大。另外，一个具有高出度且入度为 0 的用户很可能是一个恶意的信息发布者。因此，可以采用如下公式来量化用户的影响力：

$$\text{Influence}(i)=\tanh(D_O(i))\times(\alpha D_O(i)+\beta C_O(i)) \tag{4-11}$$

式中，α、β 为加权值，采用 AHP 方法计算 α、β 值分别为 0.75、0.25；tanh（D_{O}（i））表示出度的双曲正切值，表示当出度等于或接近 0 时，用户最终的影响力值等于或接近于 0。

采用式（4-11）分别计算每个时间窗口 T_k 内的节点影响力，选取前 n 个影响力值最大的用户组成集合，并对每个时间窗口内的结果进行匹配，即可跟踪和识别出随时间演变的网络论坛意见领袖。

4.3.3　算法验证

下面通过实验数据对网络论坛意见领袖识别算法性能进行测试和验证。

1．实验数据集

实验数据来源于新浪网财经论坛（http://club.finance.sina.com.cn）的技术交流版块，通过网络爬虫工具获取了 2011 年 4 月至 10 月间的发帖数据。该时间区间共有 5 128 个帖子和 421 个参与发帖的用户。按照以下规则建立论坛网络图。

（1）如果发帖人对自己所发的帖子进行回复，则不建立节点的自我指向边；

（2）如果发帖人的帖子无回帖或只有自己回复，则删除该节点；

（3）如果回帖人 B 对发帖人 A 的帖子进行了回复，则认为 B 对 A 施加了一个影响。即在两个节点间建立由 B 指向 A 的边，边的权值根据回复的次数而定。

2．算法有效性验证

为了验证算法的有效性，将算法所获取到的结果与静态网络图（以连续 5 个月为时间窗口）中所获取到的结果进行相似度值计算和对比。

在实验中，首先以连续 5 个月为时间窗口构建论坛静态网络图，如图 4-6 所示，依据式（4-11）计算出静态网络中的前 20 个意见领袖。然后使用本算法，以 4 月 1 日为起始点，以 15 天的时间周期为时间步长，构建出具有 12 个不同时间窗口的网络图。

为了选取算法中一个最优的 TOP n，分别获取每个图中前 10 个和前 20 个潜在的意见领袖。将获取的潜在意见领袖集合分别与图 4-6 中识别出的意见领袖集合进行相似度计算，具体采用 Jaccard 系数来度量相似度，因此需要计算它们之间的 Jaccard 系数，建立对应于 TOP 10 和 TOP 20 的相似度值分布图，如图 4-7 和图 4-8 所示。

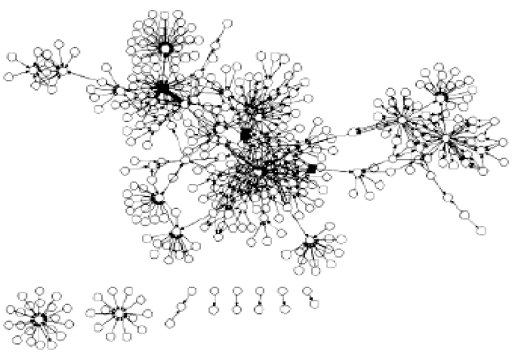


图 4-6　论坛静态网络图

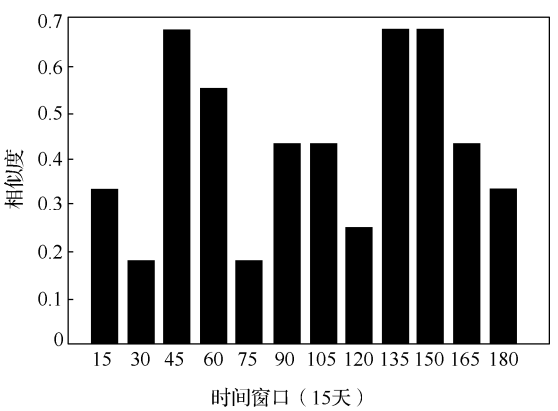


图 4-7　TOP 10 潜在意见领袖相似度值

从图 4-7 和图 4-8 可以看出，两个图的相似度值分布总体上比较相似，但图 4-7 相似度值平均要比图 4-8 高出 16%，说明采用前者的识别准确率更高。表 4-1 是结合式（4-11）和图 4-6 计算出的影响力值排名前 10 的用户列表。

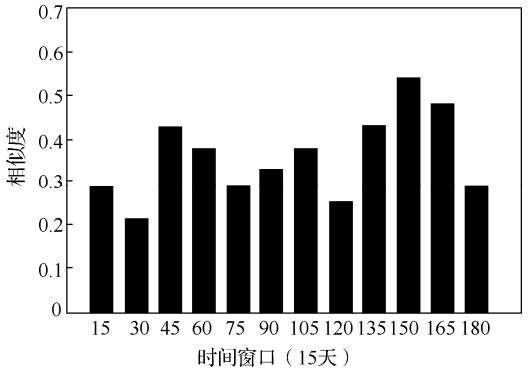


图 4-8　TOP 20 潜在意见领袖相似度值

表 4-1　影响力值排名前 10 的用户列表

用户 ID	出　　度	引入聚类	影响力值
历尽风雨见彩虹	59	0.98	44.50
财经小散	52	0.85	39.21
fcdsrgggggggg	50	0.78	37.70
frgtgt	48	0.74	36.19
云天梦	45	0.89	33.97
渐行渐远渐无言	41	0.75	30.94
俺 Q1395278391	40	0.87	30.22
飘飞的雪泥	38	0.72	28.68
2410897114hdd	37	0.72	27.93
伊凡童心	22	0.43	16.61

依据 TOP $n = 10$ 选取算法对 12 个潜在意见领袖组成的集合相互进行匹配，所获取的用户分别为：历尽风雨见彩虹、云天梦、俺 Q1395278391，并将以上 12 个集合与表 4-1 中的结果进行 Jaccard 系数计算，得到如图 4-9 所示的相似度变化图。

从图 4-9 中可以看出，不仅各个时间步长上的相似度值低于 1，相邻步长的相似度值垂直变化也比较快。

实验结果表明，根据网络论坛中用户影响力随着时间动态变化这一特性，将时间周期分割成不同的时间窗口，分别计算每个时间窗口的节点影响力，能够有效地提高意见领袖识别的准确率。

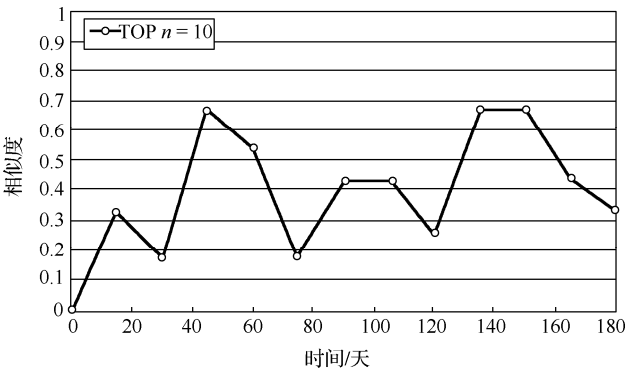


图 4-9　相似度变化图

4.4　网络水军热帖检测

由于网络论坛具有的多元化、开放性、匿名性及互动性，成为广大网民发表言论、获取信息的重要网络平台，也是网络舆情形成的主要网络平台。网络舆情包括正负两个方面，正面网络舆情是由网民发帖、点击和回帖形成的网络舆情，反映了公众对现实生活中的某些热点、焦点问题所持的具有较强影响力和倾向性的言论和观点。负面网络舆情主要是由造谣者撒布的网络谣言或者由网络水军炒作而引发的虚假网络舆情，对人们的社会生活和意识形态造成负面的影响。因此，网络水军炒作行为是引发虚假网络舆情的主要来源和推动力。

网络水军在炒作某个话题时通过发帖和回帖推动该话题迅速形成网络论坛热点话题，引起广大网民的关注，进而引发虚假网络舆情。可见，这种热点话题是由网络水军通过发帖和回帖推动的，因此称为网络水军热点话题，其帖子称为网络水军热帖。

通常，网络论坛热点话题从产生到消失需要经历一个包括潜伏期、显现期、演进期、衰退期、消解期等 5 个阶段的生命周期，如图 4-10 所示。在这些阶段，热点话题和一般话题一样，也有一个发生、发展和消失的过程，也就是从量变到质变的过程。在热点话题发生前，总会有一些征兆出现。只要及时捕捉到这些信息，加以分析处理，就能及时检测到话题幕后的推动力量，并对话题的演化过程有一个基本的认识，从而采取必要的应对措施。

网络水军对舆论的引导过程分为主题吸引和观点引导两个阶段。首先，他们将炒作的话题信息密集发布于各个网络论坛上，并通过在短时间内大量的回帖来吸引网民的眼球，引发公众围观效应。然后通过角色扮演与情感“认同”，有理有据的逻辑表达和团体协同生产等手段进一步影响网民的思想，达到左右舆论的目的。主题吸引阶段对应于热点话题生命周期中的潜伏期和演进期前期，在这一阶段，主要由网络水军不断地发帖、回帖，网民还很少参与其中。一旦进入到观点引导阶段，大量的网民参与其中，网络水军的作用就不明显了。

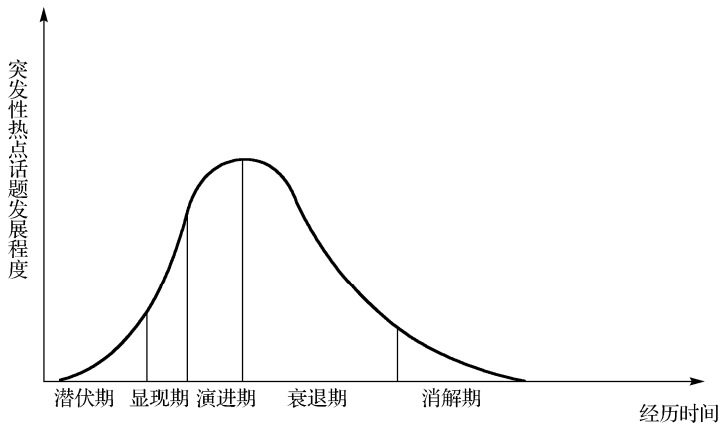


图 4-10　热点话题生命周期图

因此，网络水军热帖检测应侧重于潜伏期，首先通过分析网络水军推动的热点话题或热帖在潜伏期内的基本特征，定义并提取热点话题特征参数。然后采用机器学习算法对网络水军热帖进行分类检测，从网络论坛热点话题中准确识别出网络水军热帖。

4.4.1　热点话题特征提取

一个网络论坛热点话题可以用如下四元组来定义：

$$X=\{H_l,R_l,S_l,D_l\} \tag{4-12}$$

式中，H_l 为热点话题潜伏期回帖指数，R_l 为热点话题潜伏期新注册 ID 指数，S_l 为热点话题潜伏期简单回帖指数，D_l 为热点话题潜伏期用户 ID 离散指数。

回帖指数反映了用户针对该话题的回帖数增幅情况，定义如下：

$$H_l=\lg\frac{\sum_{i=0}^{t}p_i}{t\lambda} \tag{4-13}$$

式中，p_i 为在话题潜伏期内某一时刻 i 的回帖数，t 为话题潜伏期时间，λ 为在话题潜伏期内正常回帖率阈值。该指数考虑了由网络水军推动话题时，回帖数会呈现“指数”级的增长而非线性增长。

新注册 ID 指数反映了新注册的用户 ID 占该时期所有用户 ID 的比例，定义如下：

$$R_l=\frac{r}{R} \tag{4-14}$$

式中，r 为话题潜伏期内新注册的用户 ID 数，R 为话题潜伏期内所有用户 ID 数。

简单回帖指数反映了在话题潜伏期内内容简单的回帖所占的比例，定义如下：

$$S_l = \frac{s}{\sum_{i=1}^{t} p_i} \tag{4-15}$$

式中，s 为在话题潜伏期内简单回帖数。该指数考虑到在潜伏期有限的时间内，网络水军为使话题尽快演变成热点话题，往往注重回帖的数量而非回帖本身的质量，所以其回帖内容经常比较简单，包含的字符数也比较少。p_i 为在话题潜伏期内某一时刻 i 的回帖数。

用户 ID 离散指数反映了在话题潜伏期内同一用户 ID 发多条回帖的数量占该时期所有用户 ID 的比例，定义如下：

$$D_l = \frac{d}{R} \tag{4-16}$$

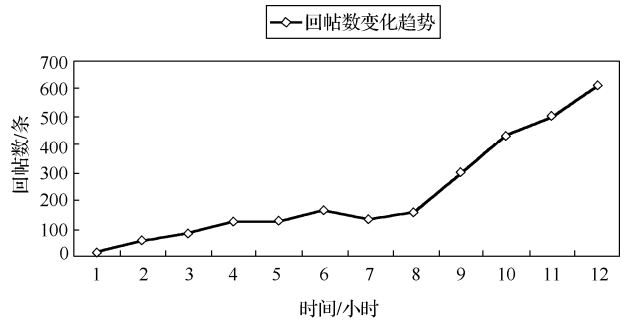
式中，d 为话题潜伏期内同一用户 ID 回帖数大于阈值 η 的个数，R 为话题潜伏期内所有用户 ID 数。该指数考虑到网络水军为使话题尽快演变成热点话题，往往在话题潜伏期内利用同一用户 ID 发表多条回帖。如果该时期同一用户 ID 的回帖数大于 η（在话题潜伏期内正常状况下同一用户 ID 回帖数阈值），则用户 ID 有网络水军注册的嫌疑。该指数还考虑到网络水军为了隐藏自己的身份，突破网络论坛对一个用户 ID“单日发帖数量限制”的约束，将会注册多个用户 ID 进行发帖、回帖。

通过对事后确认由网络推手组织、网络水军推动的热点话题的研究，发现在此类话题的演变过程中，话题热度虽然变化较大，但在回帖指数、简单回帖指数、新注册 ID 指数、用户 ID 离散指数这 4 个特征参数上均有明显的规律可循。例如，2010 年 10 月 5 日天涯社区娱乐八卦版一篇名为《感谢这样一个极品的朋友给我带来这样一个悲情的国庆》（后简称“小月月”事件）的水军帖，该帖产生后的前 12 小时内，共有 1 505 个用户 ID 参与其中，回复帖文 2 707 条。图 4-11 统计了该话题的 4 个特征参数在此时间区间上的变化情况。图中横坐标表示时间，纵坐标表示某一特征参数在对应时间区间上的大小。

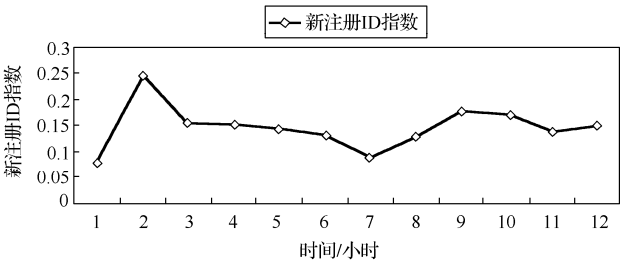
从图 4-11 可以看出：

（1）如图 4-11（a）所示，在开始的前 3～4 小时内，在网络水军的推动下，该话题的回帖数量整体处于上升趋势但变化幅度不大，随着不断演化，话题逐渐显现，从第 8 个小时开始，话题从显示期完全过渡到演进期，大量的网民参与其中，回帖数量急剧上升。

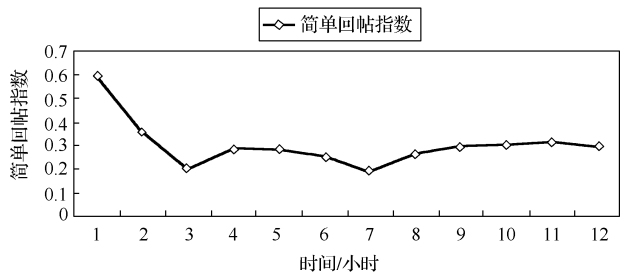
（2）如图 4-11（b）和图 4-11（c）所示，话题的新注册 ID 指数和简单回帖指数在前 2 个小时都比较高，随着时间的推移不断下降，从第 3 个小时开始，数值分别保持在 0.15 和 0.25 左右，这与由网络水军推动的热点话题所呈现的特征相吻合。处于潜伏期和显示期阶段的话题，主要是由于网络水军参与其中，网络水军通过注册新的 ID 并发送大量的简短回帖来推动话题，后期随着参与回帖的普通网民人数的不断增加，新注册用户 ID 和简单回帖数量所占的比例较潜伏期和显示期有所下降，网络水军的作用变弱。



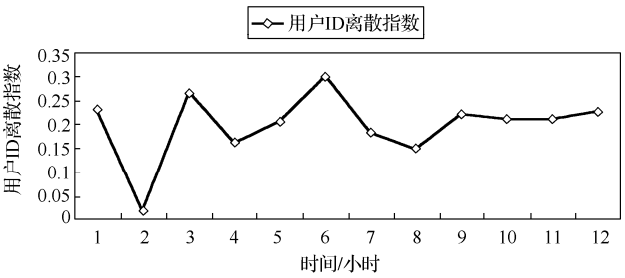
（a）回帖数变化趋势图



（b）新注册ID指数变化趋势图



（c）简单回帖指数变化趋势图



（d）用户ID离散指数变化趋势图

图 4-11　网络水军热帖统计特征

（3）如图 4-11（d）所示，随着话题过渡到演进期，大量的网民参与其中，同一用户 ID 发多条回帖的数量（这里对一个小时内发三条以上的帖子的用户 ID 进行标示）占所有用户 ID 数量的比例不仅较话题潜伏期和演进期有所降低，而且变化趋缓。

4.4.2 水军热帖检测算法

在网络水军热帖检测中，采用支持向量机（Support Vector Machine，SVM）方法。SVM 是一种经典的机器学习方法，以统计学习理论为基础，对于小样本学习问题，表现出很强的认知能力。SVM 方法的基本思想是在二维两类线性可分情况下，有很多可能的线性分类器将一组数据分割开，但是只有一个使两类的分类间隔最大。SVM 分类就是寻找一个最优分类面，尽量使该平面能够满足分类的限制条件，可以把需要分类数据集合中的所有点分开，并且尽可能地使点与该分类面距离最远。

网络水军热帖检测方法的基本思想是通过定义和提取热点话题的特征参数，利用 SVM 分类函数对热点话题幕后推动力进行分类，识别出热点话题是人为推动的还是自然形成的，如果是人为推动的，则是网络水军热帖。

网络水军热帖检测方法形式化描述如下：

给定训练集$\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$，其中 $x_i \in X$，X 表示输入空间，是所有热点话题的集合；$y_i \in Y = \{-1, 1\}$，Y 表示输出域。当热点话题由人为推动时，$y = 1$；自然形成时，$y = -1$，m 为样本数目，$1 \leqslant i \leqslant m$。

显然，网络水军热帖检测是一个二分类问题。根据上述的定义，基于 SVM 的网络水军热帖检测方法就是设计一个最优分类函数 $f(x)$: $X \rightarrow Y$，使它能够找到一个最优的线性分类面，不仅能把正常热点话题和异常热点话题分离开，还要使分类间隔最大。一个最优分类函数为：

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} a_i y_i (x \cdot x_i + b)\right) \tag{4-17}$$

整个算法实现分为两个阶段：第 1 阶段是采用 SVM 学习算法求解满足检测精度的 SVM 最优分类函数 f 和最优训练集 T；第 2 阶段是检测阶段，针对待检测热点话题样本集合 X，使用最优分类函数 f 按照顺序对集合中样本进行检测。

4.4.3 算法验证

下面通过实验数据对网络水军热帖检测算法性能进行测试和验证。

1．实验数据集

网络论坛中的热点话题一般由众多主题相似的帖子组成，并且这些帖子分布在多个论坛中，获取不同热点话题在多个论坛中的所有帖子是一件非常困难的事情。因此，本实验以单

一的网络论坛作为数据源，通过网络爬虫工具获取了网易新闻论坛（http://bbs.news.163.com）2011 年 3 月 1 日至 2011 年 5 月 1 日间的数据。这一时间区间共有 4 248 条帖子，包含 2 716 个话题，并且有超过 2 000 条帖子是孤立的。通过热点话题发现算法提取出 5 个热点话题，包含 9 条具有时间突发特性的热帖。并按以下方法对获取的数据进行处理：

（1）选取潜伏期 $t=4$ 小时，即只保留主帖自创建开始 4 小时内的回帖数据，包含回帖用户 ID、回帖时间、回帖内容、回帖用户 ID 注册时间。

（2）选取 $\lambda=50$ 条/小时，即主帖在潜伏期内的回帖率低于 50 条/小时是正常的。

（3）由于人为推动的热点话题并不需要太长的时间，所以选取新注册用户 ID 的时间阈值为 3 个月，基本包括了网络水军注册的绝大部分用户 ID，即用户 ID 注册时间少于 3 个月的视为新注册 ID。

（4）选取简单回帖内容的阈值为 10 个字符，即除去回帖内容中包含的图片、表情和标点符号后，字符个数少于 10 个的回帖视为简单回复。如果回帖内容中包含很多重复部分，重复部分少于 10 个字符，也视为简单回复。

（5）选取 $\eta=3$，即同一用户 ID 在潜伏期内对同一主帖的回帖数超过 3 条，视为该用户 ID 有网络水军注册的嫌疑。

依据以上数据处理方案，并通过式（4-13）到式（4-16）计算，可以得到如表 4-2 所示的 9 条热帖对应的量化参数。

表 4-2　热帖对应的量化参数

序　号	主帖标题	回帖指数	新注册 ID 指数	简单回帖指数	ID 离散指数
1	中石油给日本捐款 3 000 万	–1.30	0.56	0.10	0.11
2	为药家鑫的判决赌一把	–0.28	0.27	0.03	0.17
3	你做法官，如何判药家鑫？	–0.75	0.17	0.14	0.13
4	拒绝死刑，挽救家鑫，现征集万民签名	–1.60	0.67	0.10	0.10
5	新浪微博无辜封杀一剑传媒草根团队微博	0.41	0.69	0.38	0.39
6	蹲监“被死亡”，千万资产乡领导贱卖据已有	–1.35	0	0.11	0.33
7	欢呼吧！药被判死刑！	–0.40	0.14	0.03	0.57
8	震惊！副乡长抢夺民企谁汗颜？	–1.07	1	0.06	0
9	我理解地平线网友对药家鑫案的观点	–0.61	0.2	0.04	0.50

由于实际网络论坛中热点话题幕后推动力存在异常（网络水军推动）的概率低，数量少，异常的规模难以刻画，例如在表 4-2 的 9 条热帖中，只有 5 号主帖存在明显的人为推动迹象。因此实验数据是由自然形成的热点话题与网络水军推动的热点话题构成的合成话题。同时，为了更好地考察 SVM 主动学习算法的泛化能力，使实验数据集保持一定的规模，在上述 9 条热帖的基础上人为加入“小月月”事件、“凤姐”事件、“封杀王老吉”事件等 7 条已被证明是由网络水军推动的热帖，使样本集 X 中的正负样本比例为 1∶1。网络水军热帖量化参数如表 4-3 所示。

表 4-3　网络水军热帖量化参数

序号	事件标题	回帖指数	新注册 ID 指数	简单回帖指数	ID 离散指数	出　处
1	“小月月”事件	0.15	0.15	0.49	0.15	天涯社区
2	“凤姐”事件	–0.11	0.24	0.46	0.05	天涯社区
3	封杀“王老吉”事件	0.81	0.30	0.51	0.04	天涯社区
4	“康师傅”水源门事件	0.21	0.27	0.38	0.23	天涯社区
5	“犀利哥”事件	0.49	0.33	0.22	0.32	天涯社区
6	“奥巴马女郎”事件	0.64	0.36	0.35	0.16	猫扑社区
7	“贾君鹏”事件	1.16	0.60	0.79	0.42	魔兽世界吧

2．算法性能验证

在保证表 4-2 中实验数据不变的情况下，分别利用 SVM 方法与综合指标方法进行热点话题分类，评价指标是准确率、召回率和 F_1 值，两种方法的检测性能对比如图 4-12 所示，从图 4-12 中可以看出，相比于综合指标法，SVM 方法的检测性能更高，平均准确率达到 80%以上。

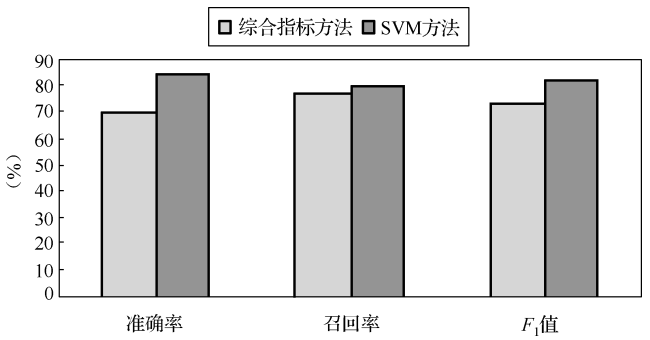


图 4-12　SVM 方法和综合指标方法检测性能对比

4.5　网络水军账号检测

网络水军受雇于网络公关公司，通过为他人发帖、回帖、造势来获得报酬，他们利用大众惯用的沟通方法在网络论坛、社交网站等平台上以聊天方式为个人或公司进行广告宣传或网络炒作，通过文章和评论来试图达到制造、影响和引导网络舆论的目的。

在网络炒作活动中，通常包括三类主体：客户、网络公关公司和网络水军，网络公关公司收到客户委托后，作为任务下发给网络推手，网络推手组织网络水军完成其任务。这样，网络公关公司、网络推手、网络水军就形成了灰色利益链，他们在实现客户目标的同时也获得了自身利益。

根据客户目标的不同，网络水军的任务一般分为两类：广告宣传和网络炒作。第一类任务是通过增加指定内容的可见率达到广告宣传的目的；第二类任务则是通过炮制网络热点，吸引广大网民围观和讨论，达到网络炒作的目的。为了完成第一类任务，网络水军需要以最快速度在各种尚没有出现该信息的网络论坛以主帖的形式发表指定内容，使其在最短时间内扩散。为了完成第二类任务，网络水军则需要短时间内在各大网络论坛大量发帖、回帖，使炒作对象在网络论坛长时间处于显眼位置，吸引网民关注，引发讨论，形成网络热点。为了完成网络炒作任务，网络水军会在全国各大论坛注册多个账号（也称为网络马甲），以不同身份登录论坛，针对论坛上若干主帖在短时间内大量回帖，达到网络炒作的目的。

下面介绍一种网络炒作的网络水军账号检测算法。

4.5.1　检测算法

1．算法基本思想

算法采用“层层逼近，逐步求精”的策略，采用人类行为统计分析、社会网络结构分析、时间特征分析等方法逐步排除正常用户和数据，不断缩小计算范围，最终确定网络水军账号。

算法流程如下：

（1）首先采用人类行为统计分析方法，统计论坛单日回帖数、日人均回帖数和日帖均回复数，将不可能发生网络炒作的时段排除，提炼出可疑区间，只对可疑区间做下一步分析，缩小分析范围；

（2）然后采用社会网络结构分析方法，对可疑区间构建单日用户协作网络，排除没有发生大规模用户协作现象的时段，提炼出高可疑数据，只对高可疑数据做下一步分析，进一步缩小分析范围；

（3）最后采用时间特征分析方法，对高可疑数据的用户回复行为时间特性进行分析，最终判定是否为网络水军。

2．论坛可疑时段识别

网络论坛的单日回帖数服从幂律分布，即大部分时间的论坛单日回帖数很小，而少数日子论坛单日回帖数很大。为了制造轰动效应，达到网络炒作的目的，网络水军通常使用多个账号在短时间内针对网络论坛上若干主帖大量地回帖，导致网络论坛的单日回帖数、日人均回帖数和日帖均回复数明显增大。因此，可以将网络论坛的单日回帖数、日人均回帖数和日帖均回复数作为识别可疑时段的指标，如果某个时段的这三项指标都大于均值，则确定该时段为可疑时段。

（1）论坛单日回帖数。该指标定义为论坛 t 日提交的回帖数之和，记作 RN_t，则有：

$$\mathrm{RN}_t = \sum_{u \in U_t} \mathrm{RN}_u^t \tag{4-18}$$

式中，N_t 为 t 日提交过回帖的用户集合，RN_u^t 为用户 u 在 t 日的回帖数。将单日回帖数大于等于均值的时段记作 S_1，则有：

$$S_t = \left[t, RN_t \geqslant \frac{\sum_{t \in T} RN_t}{|T|} \right] \tag{4-19}$$

式中，T为数据集涵盖的时段，$|T|$为数据集包含的天数，下同。

（2）论坛日人均回帖数。该指标定义为论坛 t 日回帖数与当天提交过回帖的用户数之比，记作 ARNU_t，则有：

$$\mathrm{ARNU}_t = \frac{\mathrm{RN}_t}{|U_t|} \tag{4-20}$$

将日人均回帖数大于等于均值的时段记作 S_2，则有：

$$S_2 = \left[t, \mathrm{ARNU}_t \geqslant \frac{\sum_{t \in T} \mathrm{ARNU}_t}{|T|} \right] \tag{4-21}$$

（3）论坛日帖均回复数。该指标定义为论坛 t 日回复数与当天被回复过的主帖数之比，记作 ARNP_t，则有：

$$\mathrm{ARNP}_t = \frac{\mathrm{RN}_t}{|P_t|} \tag{4-22}$$

式中，P_t 为当天被回复过的不同主帖的集合，将日帖均回复数大于等于均值的时段记作 S_3 ，则有：

$$S_3 = \left[t, \mathrm{ARNP}_t \geqslant \frac{\sum_{t \in T} \mathrm{ARNP}_t}{|T|} \right] \tag{4-23}$$

将单日回帖数、日人均回帖数、日帖均回复数均大于均值的时段定义为论坛可疑时段，记作 S，则有：

$$S = S_1 \cap S_2 \cap S_3 \tag{4-24}$$

3．用户单日回复模式分析

排除了不可能发生网络炒作的时段后，采用构建用户协作网络的方法对可疑时段的用户单日回复模式进行分析。

1）用户协作性描述

为达到网络炒作的目的，网络水军必定会使用多个账号短时间内针对同一个或几个主帖大量回帖，导致这些用户在行为上表现出很高的协作性。

为了便于描述用户的这种协作性，可以使用“用户-主帖”网络模型来表示。该网络包含两种类型的节点：用户和主帖，这里的用户表示论坛中的一个账号，主帖表示用户为了发起新的话题而发表的帖子，也称为根帖；将用户针对主帖发表的回复帖称为回帖。图 4-13（a）是 1 个包含 6 个用户、8 个主帖的“用户-主帖”网络模型，图中圆圈表示用户，正方形表示主帖，用户和主帖之间的连边表示回复关系，例如，用户 a 和主帖 2 之间的连边表示用户 a 回复过主帖 2。

用户 a 的邻节点集合定义为与节点 a 相邻的主帖节点集合，即用户 a 回复过的主帖集合，记作 Γa。用户 a 和用户 b 的协作性定义为用户 a 和用户 b 的邻节点集合的相似度，相似度用 Jaccard 系数来度量，即：

$$S_{a,b}=\frac{|\Gamma_a \cap \Gamma_b|}{|\Gamma_a \cup \Gamma_b|} \tag{4-25}$$

式中，Γ_a 和 Γ_b 分别表示用户 a 和用户 b 的邻节点集合。很明显，对于任意 a 和 b，都有 $S_{a,b}=S_{b,a}$，且 $0 \leqslant S_{a,b} \leqslant 1$。

2）用户协作网络构建

论坛用户回复行为随机性大，具有很高的异质性。如果两个或多个用户表现出很高的协作性，则有理由怀疑其为网络水军账号。在“用户-主帖”网络模型的基础上，构建单日用户协作网络，对该网络的聚类特性进行分析，确定高可疑时段。

用户协作网络的构建方法如下：将用户抽象为节点，如果两个用户的协作性大于 0，即他们均回复过至少同一个主帖，则在这两个用户之间建立连边，边的权值定义为两个用户的协作性。图 4-13（b）是根据图 4-13（a）构建的用户协作网络。从图 4-13（b）可以看出，用户 a、b 和 c 之间的协作性为 1，即他们的回复对象完全相同，高度可疑。

为了更清楚地观察节点间的协作性，快速确定高可疑用户，按照边的权值对用户协作网络进行删减，仅保留其协作性大于一定阈值的边。如果仅保留图 4-13（b）中权值大于 1/3 的边，则得到图 4-13（c）。协作性高的用户表现出明显的社团特性，将此类用户看作高可疑用户。

4．高可疑用户回复行为

人类打电话行为在时间上具有一定的规律性，工作时段活跃度高，休息时段活跃度低，网民回帖行为也具有类似特性。因此采用时间特征分析方法，对用户回帖行为时间特征进行分析，判定某天是否发生了网络炒作。对于确定发生了网络炒作的时段，根据网络水军相互协同这一特征推断出以“簇”形式出现的论坛用户即为网络水军账号，实施同一网络炒作的水军账号形成了水军军团，同一簇内用户共同回复的话题即为网络炒作的内容。

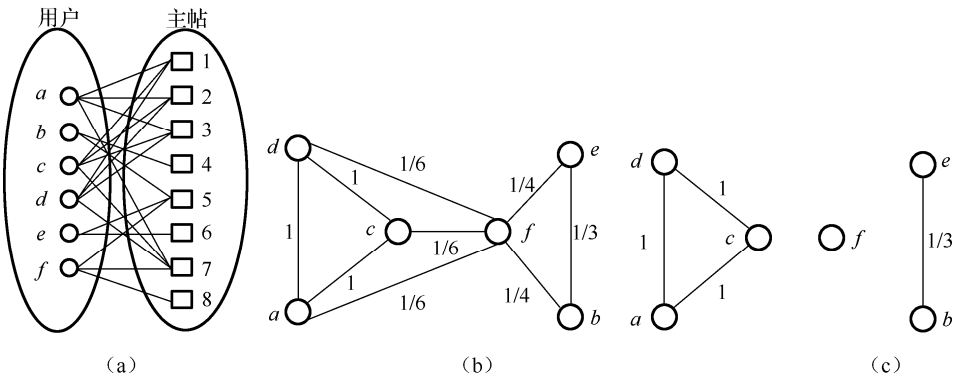


图 4-13 “用户-主帖”网络和用户协作网络模型示例

4.5.2 算法验证

下面通过实验数据对网络水军账号检测算法性能进行测试和验证。

1．实验数据集

实验数据来源于“新浪网－娱乐论坛－影视世界版块－影行天下子版块”2010 全年的发帖、回帖和用户信息。用 post、reply 和 user 三个数据表来存储采集到的数据，其中 post 表存储主帖信息，包括主帖 ID、发帖时间、发帖用户 ID、标题、内容；reply 表存储回帖信息，包括回帖用户 ID、回帖时间、回帖内容、对应主帖 ID；user 表存储相关用户信息，包括用户 ID、用户名、用户级别、在线时间、注册时间。

数据集共包含 4 407 个主帖、80 990 个回帖和 13 099 个用户，其中发表过主帖的用户 1 011 个，发表过回帖的用户 12 929 个，2012 年全年没有发帖或回帖的用户排除在外。

2．算法有效性验证

1）可疑时段计算

按照式（4-18）到式（4-24）对数据集进行统计分析，并计算三项统计指标的最小值、最大值及均值，计算结果如表 4-4 所示。

表 4-4 三项统计指标的计算结果

指　标	统　计　量			
	Min	Max	Avg	>A
RN	7	18 824	221	69
ARNU	1	29.41	2.15	103
ARNP	1	896.38	9.85	58

由表 4-4 可知，三项统计指标的异质性均非常强，大多数日子的取值都比较小。统计发

现单日回帖数不小于均值的共 69 天，单日人均回帖数不小于均值的共 103 天，单日帖均回复数不小于均值的共 58 天，同时满足三个条件的共 45 天，即为可疑时段 *S*。

2）高可疑时段确定

通过逐天分析可疑时段的用户回复模式，发现有 29 天的用户协作网络发生了明显聚类现象，将其确定为高可疑时段。图 4-14 显示了其中 4 天的用户协作网络，从图 4-14 可以看出，4 天用户回复行为均表现出极高的协作性。图 4-14（b）显示了 12 月 3 日仅保留权值大于 0.9 边的用户协作网络，除零星用户处于离散状态外，其他用户都聚集成为 8 个簇，同一簇内的用户协作性高达 0.9，即回复对象非常接近，高度可疑。

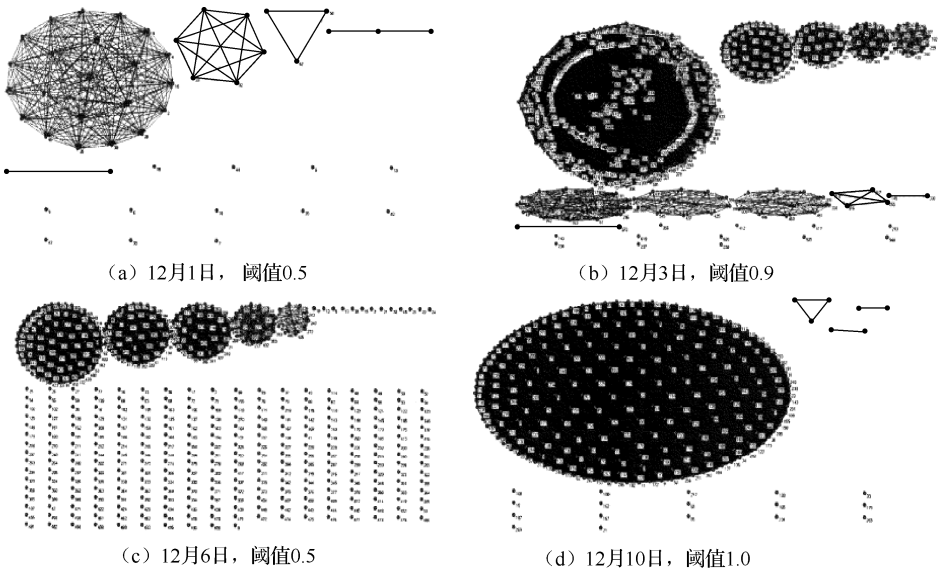


图 4-14　高可疑时段用户协作网络示例

3）网络水军账号确定

为了确认高度可疑的 29 天中形成簇的用户是否为网络水军，逐天分析这些用户的回帖时间分布，统计分析结果发现，其中 7 天的用户回帖时间分布严重偏离正常用户的回帖时间分布，由此可断定这 7 天的网络论坛发生了网络炒作，它们分别是 12 月 2 日、12 月 3 日、12 月 5 日、12 月 6 日、12 月 10 日、12 月 12 日和 12 月 13 日。图 4-15 是用户回帖时间模式比较，显示了 2010 年全年及 12 月 3 日、12 月 6 日和 12 月 10 日的回帖时间在一天中的分布，其中横坐标为时间，纵坐标为该段时间的回帖数。为了便于显示，将 12 月 3 日、12 月 6 日和 12 月 10 日的统计数据分别扩大 2 倍、10 倍和 10 倍。

从图 4-15 可以看出，2010 年全年的零点回帖数较低，之后逐渐下降，并在 7 点达到谷底，这段时间正好对应人们的休息时间。之后回帖数快速上升，9 点至 23 点之间回帖数都

保持在 3 500 以上，其中 9 点到 18 点的回帖数略高于 18 点之后。统计结果与人们的作息规律非常吻合，也与人类打电话时间模式相一致。

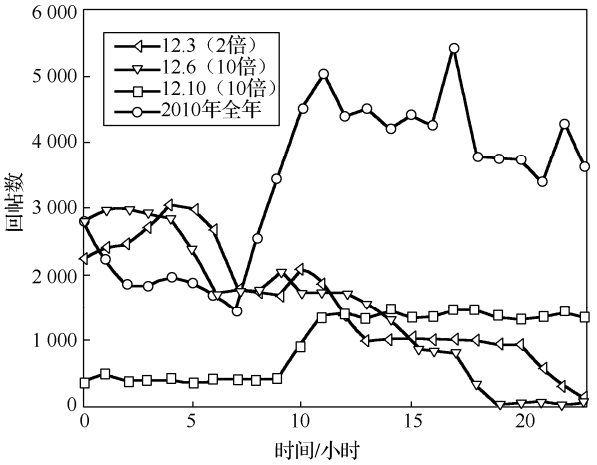


图 4-15　用户回帖时间模式比较

观察 12 月 3 日的回帖模式，发现零点回帖数很大，之后的 5 个小时持续攀升，并在 4 点和 5 点达到最高峰；之后快速下降，9 点至 12 点回帖数均低于当天零点；13 点至 20 点，回帖数稳定在 500 左右，不到零点时的一半；之后继续下降，直到 23 点回帖量达到最低值。可以看出，12 月 3 日的用户回帖时间分布与人类作息时间完全违背。12 月 6 日的回帖时间分布与 12 月 3 日几乎相同，12 月 10 日的回帖模式与 12 月 3 日、12 月 6 日虽然不同，但表现出异乎寻常的稳定性，也不符合人类作息规律。采用同样方法分析另外 4 天的用户回帖时间模式，发现其用户回帖时间模式也明显偏离正常用户行为特征。

通过对发生网络炒作的 7 天的用户协作网络的统计分析，发现簇内共包含不同账号 556 个，构成了 1 个网络水军军团，炒作内容为当时即将上映的电影《赵氏孤儿》。

采用人工分析方式，对算法检测出的网络水军账号逐个进行分析，发现它们均为网络水军账号，算法的准确率达 100%。除算法发现的网络水军账号外，没有发现其他的可疑账号，因此该算法的漏报率为 0。

综上所述，该算法将人类行为统计分析、社会网络分析和时间特征分析等方法结合起来，逐步排除正常用户和数据，不断缩小计算范围，最终识别出网络水军账号，具有准确率高、计算量小、运算速度快等特点。

第5章

话题检测与跟踪技术

5.1 引言

话题检测与跟踪（Topic Detection and Tracking，TDT）是自然语言处理和信息检索领域提出的研究课题，最初是由美国国防高级研究计划署（DARPA）发起研究的，旨在没有人工干预的情况下自动检索、判断和识别新闻数据流中的话题。从 1998 年开始，在 DARPA 的支持下，美国国家标准技术研究所（NIST）每年都举行 TDT 测评会议，发表有代表性的成果和论文。由于每次 TDT 测评会议都对 TDT 研究子课题提出了新的设想与方向，因此相应的测评任务也随之做相应的调整。截至最后一次 TDT 测评会议（TDT2004），TDT 测评会议共设立了 6 项测评任务，分别为：新事件检测（New Event Detection）、报道关系检测（Story Link Detection）、话题检测（Topic Detection）、话题跟踪（Topic Tracking）、自适应话题跟踪（Adaptive Topic Tracking）和层次话题检测（Hierarchical Topic Detection），其中话题检测与跟踪是核心问题。

国外对 TDT 技术的研究起步较早，取得了很多的成果，这些成果成为后来研究 TDT 技术的重要参考文献。TDT 技术的最初应用主要是新闻出版领域，用于新闻流的话题检测和事件跟踪。后来被扩展到互联网上，用于检测和跟踪以话题词为中心的互联网新闻热点话题以及流行词。

国内在这方面的研究开展得要晚一些，研究的重点是面向中文的 TDT 技术，并在中文信息处理和检索中得到了广泛的应用。

TDT 是从一篇文章的主题作为切入点，通过对文章主题的发现和跟踪，把各种分散的信息进行有效的汇集，并且组织成线索提供给用户进行查阅，厘清一个主题事件的来龙去脉，把握整个事件的整体和细节。例如，在网络舆情监测中，通过 TDT 技术对各种信息源的监测和分析，从中识别出各种突发事件以及事件的演化过程。TDT 技术还可以应用于证券市场分析等领域，用途比较广泛。

本章主要介绍 TDT 的基本概念、相关技术、话题检测算法、话题跟踪算法、热点话题检测等内容。

5.2 基本概念

5.2.1 TDT 目标和任务

1．相关术语

为了更好地理解 TDT 技术，下面介绍 TDT 中常用的术语。

（1）话题（Topic）：在 TDT 中，话题通常被定义为与现实世界中某个事件相关的新闻故事的集合。在最初的研究中，话题和事件含义基本相同。一个话题是指因某些原因、条件引起，发生在特定时间和地点，有一定的参与者或涉及者，并可能伴随某些必然结果的一个事件。例如，“2001 年 9 月 11 日美国世贸大厦遭受到恐怖袭击”。后来的话题概念要相对宽泛一些，它包括一个核心事件或活动以及所有与之直接相关的事件和活动。如果一篇报道讨论了与某个话题的核心事件直接相关的事件或活动，那么也认为该报道与此话题相关。例如，搜救 9·11 事件的幸存者、安葬死难者等都被看作与“9·11 恐怖袭击”这个话题相关。

（2）事件（Event）：事件通常是指在特定时间和地点发生的事情。可以简单地认为话题就是若干对某个事件相关报道的集合。例如，“2001 年 9 月 11 日美国世贸大厦遭受到恐怖袭击”是一个事件而不是话题，而“美国世贸大厦遭受到恐怖袭击”是话题而不是事件，事件是话题的实例，与一定的活动相关。

（3）故事（Story）：故事是对某个事件的相关报道。在 TDT 中，通常是指一个与话题紧密相关的、包含两个或多个独立陈述某个事件的子句的新闻段落。

（4）主题（Subject）：主题的含义更广一些，主题可以涵盖多个类似的具体事件或者根本不涉及任何具体事件，而话题则与某个具体事件相关。例如，“恐怖袭击”是一个主题，而“美国世贸大厦遭受到恐怖袭击”则是一个话题。又如，“自然灾害”是一个主题，而属于该主题类别的文本未必有与之直接相关的事件发生，例如讲述自然灾害预防的文章等。

2．TDT 目标

TDT 测评会议对参评的 TDT 系统设定的目标是实现一个功能强大、用途广泛的全自动算法，用以判断自然语言数据的话题结构。换句话说，TDT 的目标把连续文本分割成一系列相互独立的故事（Story），监测以前从未出现过的故事，并把若干故事组合成一个新闻话题。

TDT 的最初目标是创造一种核心技术，可以监测各种渠道的新闻广播，发现世界上发生的一些新奇有趣的消息，以供进一步研究。

事实上，研究者对新闻报道中的新信息很感兴趣，但是又没有一种高效的方法，能够

有效地处理每天产生的大量信息。这种需求推动了 TDT 技术的研究，为了能够跟踪具体的事件，不仅需要跟踪到更大范围的“题目”，还需要更深入地理解和利用文章的含义。因此需要结合自然语言处理、信息检索等相关知识来构造模型，尽可能地利用文章中现有信息来合理地表示该事件。

后来，人们又提出了很多 TDT 技术的应用目标，如个性化推荐等，希望通过不同的人群对不同文章兴趣度的区别，能够自动跟踪不同个体所关心的主题或者事件，在一篇新文章中出现他所关注的事件或主题时，能够自动通知本人，达到自动推荐的目的。

3．TDT 任务

TDT 是一种交叉性技术，以自然语言处理技术作为主要技术支撑。因此 TDT 测评会议对 TDT 任务进行了细化，根据不同的应用需求，将 TDT 任务分成 5 个子任务。

（1）新闻报道切分：该任务的目标是将一个语言信息流分割为不同的独立新闻报道。由于文本信息流本身就是以单个报道形式出现的，不存在切分问题。因此该任务只适用于对来自广播、电视等媒体的音频数据处理。一段新闻节目通常包含了很多条报道，这些节目本身很少在不同的新闻报道间设置明显的分隔标记，例如商业广告很可能出现在某一报道的中间。要切分的语料或数据可以是音频记录本身，也可以是从音频记录得到的文字记录。

（2）新事件检测：该任务的目标是从新闻报道信息流中检测出对一个新话题的首次报道，该任务也被看作对一个话题检测系统的透明测试。在新事件检测中常用的典型方法是采用向量或概率分布形式的特征项集合来代表每篇报道，每遇到新来的报道，就将其特征项集合与过去所有报道的特征项集合进行相似度比较，以此来判断该报道是否描述了一个新的话题。

（3）报道关系检测：该任务的目标是判断两个随机选择的新闻报道是否讨论同一个话题。与其他任务相比，该任务没有直接的应用目标，因此对该任务的研究并没有受到重视。在报道关系检测中，需要使用某种相似度度量方法来计算报道关系的相似度，经过测试表明，余弦系数方法是比较有效的。

（4）话题检测：该任务的目标是检测出未知的话题以及相关报道。话题检测关注的是将某个话题的所有报道归入一个话题类，因此它是一个无监督学习的聚类问题。通常的聚类可以看作基于全局信息的聚类，即在整个数据集合上进行聚类，而话题检测中的聚类是增量式的，即做出最终的决策前，不能或只能看到前面有限数量的报道。在话题检测研究中，大多数采用传统的自然语言处理方法，如中心向量法、*K* 最近邻居（*K*-Nearest Neighbor，KNN）、*K*-均值（*K*-MEANS）、单遍聚类算法等。

（5）话题跟踪：该任务的目标是给出某一话题的一组样本报道，通过训练得到话题模型，然后在后续报道中找出所有讨论目标话题的报道。该任务的实质是通过有监督的学习过程，利用非常少的正例数据和大量的反例数据来获得一个分类器，用于区分新报道与话题的

相关性，因此可以把话题跟踪看作一种特殊的二值分类问题。在话题跟踪的研究中，常用的分类方法有 KNN、决策树、Rocchio 算法、SVM、语言模型算法、概率模型、基于查询的方法等，其中比较有效的方法是 KNN 算法以及多种算法的组合。

5.2.2 TDT 语料

语料可以看作用于测试 TDT 系统或算法有效性的数据集合。TDT 语料是选自大量新闻媒体的多语言新闻报道集合。TDT 语料有 5 期，分别是 TDT-Pilot、TDT2、TDT3、TDT4 和 TDT5，其中 TDT5 只包含文本形式的新闻报道，而其他语料同时包含文本和广播两种形式的新闻报道。

TDT 测评最早使用的语料是 TDT-Pilot（TDT Pilot corpus），TDT-Pilot 收集了 1994 年 7 月 1 日到 1995 年 6 月 30 日之间约 16 000 篇新闻报道，主要来自路透社新闻专线和 CNN 新闻广播的翻录文本，TDT-Pilot 标注过程没有涉及话题的定义，而是由标注人员从所有语料中选择各个领域的 25 个事件作为话题检测与跟踪的对象。TDT2 收集了 1998 年前 6 个月的英文、中文两种语言形式的新闻报道，人工标注了 200 个英文话题和 20 个中文话题。TDT3 收集了 1998 年 10 月到 12 月英文、中文和阿拉伯文三种语言的新闻报道，人工标注了 120 个英文和中文话题，并选择部分话题用阿拉伯文进行了标注。TDT4 收集了 2000 年 10 月到 2001 年 1 月英文、中文和阿拉伯文三种语言的新闻报道，分别采用三种语言对 80 个话题进行了人工标注。TDT5 收集了 2003 年 4 月到 9 月的英文、中文和阿拉伯文三种语言的新闻报道，人工标注了 250 个话题，其中 25%的话题同时具有三种语言的表示形式，其他话题则以相同的比例均匀地分配给三种语言分别进行了标注。此外，TDT5 中每种语言的话题均来自于该语言当地媒体的报道。

在所有 TDT 语料中，还对报道与话题的相关性进行了标注，其中，TDT2 和 TDT3 采用“YES”、“BRIEF”和“NO”三种标识对报道与话题的相关性进行标注。当报道的内容与话题绝对相关时标注为“YES”，而报道与话题相关的内容低于 10%时则标注为“BRIEF”，不相关则标注为“NO”。而 TDT4 与 TDT5 只采用相关“YES”和不相关“NO”两种标识来标注报道与话题的相关性，其中，相关报道不仅需要相关于话题的核心内容，同时还需要包含话题的部分信息。但是，报道与话题的相关性并没有 TDT2 和 TDT3 中所要求的长短（BRIEF）之分，只要存在相关信息都被标注为“YES”。

5.2.3 TDT 评价指标

在 TDT 的评价标准中，采用准确率、召回率、漏报率和误报率 4 个评价指标来评价被测 TDT 系统的性能，各项评价指标定义如下：

（1）准确率（P）：系统正确识别出的关于某一话题的报道数量与所有识别出的报道总数之比，也称为查准率，计算公式为 $P=A/(A+B)$，其中，A 为系统正确识别出的相关报道数

量，B 为系统将不相关报道错误判断为相关报道的数量。

（2）召回率（R）：系统正确识别出的关于某一话题的报道数量与语料库中描述该话题的报道总数之比，也称为查全率，计算公式为 $R=A/(A+C)$，其中，A 为系统正确识别出的相关报道数量，C 为系统未识别出的相关报道数量。

（3）漏报率（M）：系统没有识别出的关于某一话题的报道数量与语料库中描述该话题的报道总数之比，计算公式为 $M=C/(A+C)$，其中，A 为系统识别出的相关报道数量，C 为系统未识别出的相关报道数量。

（4）误报率（F）：系统将某一话题不相关报道错误判断为相关报道的数量与语料库中没有描述该话题的报道总数之比，计算公式为 $F=B/(B+D)$，其中，B 为系统将不相关报道错误判断为相关报道的数量，D 为系统未识别的不相关报道数量。

在话题检测与跟踪中，对一个 TDT 系统的性能评价还使用了归一化识别代价$(C_{\mathrm{Det}})_{\mathrm{Norm}}$指标，它由系统的漏报率和误报率计算得到，计算公式如下：

$$(C_{\mathrm{Det}})_{\mathrm{Norm}}=\frac{C_{\mathrm{Det}}}{\min(C_{\mathrm{Miss}}\times P_{\mathrm{Target}},C_{\mathrm{FA}}\times P_{\mathrm{Nontarget}})} \tag{5-1}$$

$$C_{\mathrm{Det}}=C_{\mathrm{Miss}}\times P_{\mathrm{Miss}}\times P_{\mathrm{Target}}+C_{\mathrm{FA}}\times P_{\mathrm{FA}}\times P_{\mathrm{Nontarget}} \tag{5-2}$$

其中：

（1）C_{Det} 为系统的错误识别代价，它又由式（5-2）计算得到。

（2）C_{Miss}、C_{FA} 分别为漏报和误报的代价，它们的值通常根据应用预先给定。在大多数 TDT 测评任务中，它们分别取 10 和 1，即认为漏报的代价要高得多。

（3）P_{Miss}、P_{FA} 分别为系统识别的漏报率和误报率，它们可由系统输出与标准答案对照的结果计算得到，其计算公式为 P_{Miss} = 漏检数量/目标数量、P_{FA} = 误报数量/非目标数量。

（4）P_{Target} 为一个先验的目标出现概率，即 $P_{\mathrm{Nontarget}}=1-P_{\mathrm{Target}}$，表示关于某个话题的新闻报道出现的可能性，它的值通常也是根据具体应用给出。

为了使所得到的性能指标落在更有意义的范围内，将错误识别代价 C_{Det} 做归一化处理得到$(C_{\mathrm{Det}})_{\mathrm{Norm}}$。在式（5-1）中，分母部分实际上是一个最小的预期代价，它是由系统对每一项识别给出的全部肯定猜测或全部否定猜测而得到的。归一化处理后的识别代价的最小值为 0，表示系统性能最佳，最大值为 1，表示系统性能较差。

在评价一个 TDT 系统性能时，通常采用两种计算方法：基于报道加权和基于话题加权。在几次 TDT 测评会议上只采用基于话题加权的计算方法，因为这种方法能够保证系统识别能力不会受到某些话题包含了大量报道的影响。

除了归一化识别代价指标外，还使用识别错误权衡曲线来直观地刻画漏报率与误报率之间的反比关系，它是根据系统对每个判断给出的可能性大小来绘制的。

5.3 相关技术

在网络舆情分析中，应用 TDT 技术对海量的文本信息按照话题进行归类和组织，并对特定的话题进行跟踪，以取代人工来完成话题检测与跟踪任务，使用户能够在动态环境下发现热点话题和舆情，并跟踪舆情的变化趋势。

在话题检测和跟踪中，需要解决以下几个问题：

（1）报道和话题的表示模型；

（2）特征项权重计算；

（3）话题和报道间的相似度计算；

（4）文本分类与聚类的策略选择。

下面介绍解决这些问题的常用方法，这些方法的有效性在 TDT 测评中得到了验证，并且已被广泛应用于实际中。

5.3.1 表示模型

为了判断一个报道是否与某个话题相关，首先需要使用适当的模型来表示报道和话题，以便对两者的相关度进行计算和比较。常用的表示模型有向量空间模型和语言模型。向量空间模型介绍见 2.5.1 节，下面简单介绍语言模型。

语言模型是一种概率模型，假设报道中出现的词 δ_n 各不相关，则某个报道 S 与话题 C 相关的概率为：

$$P(C|\ S)=\frac{P(C)P(S|\ C)}{P(S)}\approx P(C)\prod_n\frac{P(\delta_n|\ C)}{P(\delta_n)} \tag{5-3}$$

式中，$P(C)$为任何一个新报道和话题相关的先验概率，$P(\delta_n|C)$是表示词 δ_n 在话题 C 中的生成概率，$P(\delta_n|C)$可以表示成一个两态的混合模型，如图 5-1 所示。

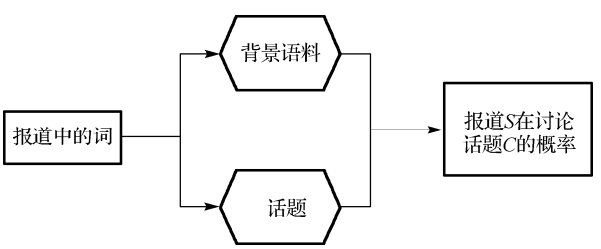


图 5-1 $P(\delta_n|C)$混合模型示意图

在两态的混合模型中，一个状态是词在该话题中所有报道的分布，另一个状态是词在整个语料中的分布，这样就构成了一个词的生成模型。采用最大似然估计（ML）来计算该

模型中的两个状态，即该话题的所有报道中 δ_n 出现的次数除以该话题所有报道所包含的总词数。因为话题语言模型很稀疏，这里需要解决未见词的零概率问题，通常采用线性插值法来加入背景语言模型：

$$P'(\delta_n|C)=\alpha P(\delta_n|C)+(1-\alpha)P(\delta_n) \tag{5-4}$$

一般语言状态分布和话题状态分布可以采用期望最大化（EM）算法来估算，EM 算法能够对与话题相关的词汇赋予较高的概率。

在话题检测与跟踪中，人们提出了多种语言模型，如隐马尔可夫模型、指数语言模型、层次语言模型、语义模型等，其中效果较好的是 LDA（Latent Dirichlet Allocation）模型。关于 LDA 模型的详细介绍见 6.3.1 节。

5.3.2　相似度计算

对所有的主题 C_1, C_2, …, C_n，要判断某个报道 S 属于哪一个主题，就需要计算报道和各个主题之间的相似度，通过相似度值和阈值进行比较，做出最终的判断。

下面简单介绍几种常用的相似度计算方法。

1．内积

文档是由相互独立特征词 T_1, T_2, …, T_m 构成，令 $D=(D_1, D_2, \cdots, D_n)$表示由 m 个特征词构成的 n 个文档的文档集合，其中 $D_j=(d_{1j}, d_{2j}, \cdots, d_{mj})$是文档向量，$d_{ij}$ 表示特征词 i 出现在文档 j 中的权重。

设 $D_i=(d_{1i}, d_{2i}, \cdots, d_{mi})^{\mathrm{T}}$，则 D_i 与 D_j 之间的相似度用内积表示如下：

$$\mathrm{Sim}(D_i,D_j)=\sum_{k=1}^{m} d_{ki}\times d_{kj} \tag{5-5}$$

权重的选择可以使用多种方法，如果选择二值权重计算方法，即特征词 i 出现在文档 j 中，$d_{ij}=1$；如果特征词 i 不出现在文档 j 中，$d_{ij}=0$，则有 $\mathrm{Sim}(D_i, D_j)=|D_i\cap D_j|$，其中$|D_i\cap D_j|$表示同时出现在文档 D_i 和 D_j 中的特征词。

2．Dice 系数

文档 D_i、D_j 的 Dice 系数定义为：

$$\mathrm{Sim}(D_i,D_j)=\frac{2\sum_{k=1}^{m} d_{ki}\times d_{kj}}{\sum_{k=1}^{m} d_{ki}^2+\sum_{k=1}^{m} d_{kj}^2} \tag{5-6}$$

如果选择二值权重方法，则有

$$\mathrm{Sim}(D_i, D_j) = \frac{2|D_i \cap D_j|}{|D_i| + |D_j|} = \frac{2C}{A + B} \tag{5-7}$$

式中，$C = |D_i \cap D_j|$是同时出现在文档 D_i 和 D_j 中的特征词数，$A = |D_i|$和 $B = |D_j|$分别表示 D_i、D_j 特征词数。

3．Jaccard 系数

文档 D_i、D_j 的 Jaccard 系数定义为：

$$\mathrm{Sim}(D_i, D_j) = \frac{\sum_{k=1}^{m} d_{ki} \times d_{kj}}{\sum_{k=1}^{m} d_{ki}^2 + \sum_{k=1}^{m} d_{kj}^2 + \sum_{k=1}^{m} d_{ki} \times d_{kj}} \tag{5-8}$$

4．余弦系数

文档 D_i、D_j 的余弦系数定义为：

$$\mathrm{Sim}(D_i, D_j) = \frac{\sum_{k=1}^{m} d_{ki} \times d_{kj}}{\sqrt{\sum_{k=1}^{m} d_{ki}^2 \times \sum_{k=1}^{m} d_{kj}^2}} \tag{5-9}$$

5．欧几里得度量

欧几里得度量又称欧几里得距离，采用欧几里得距离来度量文档间相似度，即文档 D_i、D 的欧几里得度量定义为：

$$\mathrm{Sim}(D_i, D_j) = \sqrt{\sum_{k=1}^{m} (d_{ki} - d_{kj})^2} \tag{5-10}$$

5.3.3 特征项选取

文本的表示及其特征项选取是文本挖掘、信息检索中的一个基本问题，通过表示模型抽取文本中的特征词进行量化，使一个无结构的原始文本转化为结构化文本，这样计算机才能对文本内容进行处理和识别。

文本的表示模型有向量空间模型、语言模型等，比较常用的是向量空间模型。在向量空间模型中，通过特征项来表示文本向量中的各个维，特征项选取方法非常关键。直接用分词和词频统计方法来得到特征项，可能导致向量维度非常大，给后续处理带来很大的计算开销，而且还会影响到分类、聚类算法的性能。因此，需要对文本向量做净化处理，在保证原

文含义的基础上，找出最具代表性的文本特征项。这个问题归结为找到一种低维度的特征选取方法。

特征选取方法主要有 4 种：

（1）使用映射或变换的方法把原始特征变换为较少的新特征；

（2）从原始特征中挑选出一些最具代表性的特征；

（3）根据专家的知识挑选最有影响的特征；

（4）使用统计方法找出最具分类信息的特征，这种方法是一种比较精确的方法，人为因素的干扰较少，尤其适合于文本自动分类挖掘。

基于统计的特征选取方法通过构造评估函数，对特征集合中的每个特征进行评估和打分，这样每个词语都获得一个评估值，又称为权值。然后将所有特征按权值大小排序，提取预定数量的最优特征作为提取结果的特征子集。对于这类算法，决定文本特征提取效果的主要因素是评估函数的优劣。基于统计的特征选取方法主要有文档频率（DF）、信息增益（IG）、互信息（MI）、卡方检验（CHI）等，见 2.4.1 节。而特征权值计算方法主要采用 TF-IDF 法，见 2.5.2 节。

实验表明，CHI、IG 和 DF 的性能明显优于 MI；CHI、IG 和 DF 的性能大体相当，并且当保留的特征项超过总数的 15%后，系统的性能趋于稳定，因此三者都可以滤掉 85%以上的特征项。此外，DF 还具有算法简单、质量高的优点，可以取代 CHI 和 IG。

5.3.4　文本聚类

话题检测的任务是将关于某个话题的所有报道自动归入一个话题类，它是在事先没有分类体系和训练语料的情况下对报道进行聚类分析，因此它是一个无监督学习的聚类问题。文本聚类通常是在已有的文本集合上进行聚类分析，给出一个最佳的划分，而不需要预先对文档类别进行标注。

文本聚类分析是一种无监督的学习过程，不需要预先对文档手工标注类别，即不依赖于文档集合划分的先验知识，仅仅根据文档集合内部的文档对象彼此之间相似度关系并按照某种准则进行文档集合划分。文本聚类划分主要依据于这样的聚类假设：同类文档彼此之间的相似度较大，而不同类文档之间的相似度较小。由于文本聚类分析不需要事先定义文档类别，对获取大规模多元数据集合的结构特征是有效的，能够发现数据之间所隐含的某些关系，因此在数据挖掘和知识发现领域中得到了广泛应用。

1．文本聚类的步骤

文本聚类过程可以分为以下 3 个步骤：

（1）文本表示。把文档表示成聚类算法能够处理的形式，最常用的文本表示方法是向量空间模型。

（2）聚类算法。使用无监督学习算法对文档集合进行划分，文本聚类算法有很多种，但是没有一个通用的算法能够解决所有的聚类问题，因此需要针对所要解决问题的特点，选择合适的聚类算法。另外，聚类算法的选择往往与相似度计算方法相对应。在文本挖掘中，最常用的相似度计算方法是余弦系数法。

（3）效果评估。通常使用准确率、召回率、漏报率和误报率等评价指标对聚类效果进行评估，也是对聚类算法性能的评估。由于没有训练文档集合，因此评估算法的聚类效果往往是比较困难的。常用的评估方法是选择人工已分好类或者已做好标记的文档集合作为测试集合，当聚类结束后，将聚类结果与已有的人工分类结果进行比较来评估算法的聚类效果。

2．常用文本聚类算法

文本聚类算法有很多种，常用的算法有层次方法、划分方法、基于密度的方法、基于网格的方法、基于模型的方法等。

1）层次方法

这种方法是通过创建一个层次结构来分解给定的数据集，可以分为自上而下（分裂）和自下而上（合并）两种操作方式。为弥补分裂与合并的不足，层次合并经常需要与其他聚类方法相结合，如循环定位等。典型的方法有：

（1）BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）方法：该方法引入了两个概念：聚类特征和聚类特征树（CF 树），用于概括聚类描述。其工作过程包括两个阶段，首先利用 CF 树对对象集合进行划分，然后采用某个聚类方法对 CF 树的叶子节点进行聚类和优化。BIRCH 方法对增量或动态聚类非常有效。

（2）CURE（Clustering Using REprisentatives）方法：该方法选择基于质心和基于代表性对象方法之间的中间策略，它不用单个质心或对象来代表一个簇，而是选择数据空间中固定数目的点，一个簇的代表点按照下列方式产生：首先选择簇中分散的对象，然后根据一个特定的分数或收缩因子向聚类中心进行收缩。在算法的每一步，都要将最近距离的代表点对的两个簇进行合并。CURE 方法比较适合处理球形和相似大小的聚类问题，在孤立点处理上更加健壮。

（3）ROCK（RObust Clustering using linKs）方法：该方法通过聚集的互连性与用户定义的静态互连性模型相比较来度量两个簇的相似度。两个簇间的互连性可以用两个簇间的交叉链数目来定义，而交叉链是指两个点共同的近邻数目。换句话说，簇间相似度可以用来自不同簇而有相同近邻的点的数目来度量。该方法首先根据相似度阈值和共享近邻的概念从给定的数据相似度矩阵构建一个稀疏的图，然后在这个稀疏图上执行一个层次聚类算法。ROCK 方法适用于分类属性。

（4）CHEMALOEN：该方法也称为变色龙算法，首先通过一个图划分算法将数据对象聚类成大量相对较小的子聚类，然后用一个层次聚类算法通过反复合并子类来找到真正的结

果簇。在聚类过程中，如果两个簇间的互连性与近似性和簇内对象间的互连性与近似性高度相关，则合并这两个簇。该方法在确定最相似子类时，同时考虑了簇间和簇内的互连性和近似性，而不依赖于一个静态的模型，能够自动适应被合并的簇内的特征。CHEMALOEN 方法改进了 CURE 方法和 ROCK 方法的缺点，CURE 方法忽略了两个簇中对象的聚类互连性信息，而 ROCK 方法强调了对象间的互连性，却忽略了对象间的近似性信息。

2）划分方法

这种方法是为获得全局最优结果而穷举所有可能的对象划分。为此，大多数应用采用 1～2 种常用的启发方法，如采用 K-MEANS 方法和 K-MEDOIDS 方法。

（1）K-MEANS：该方法以 k 为参数，将 n 个对象划分为 k 个簇，使簇内具有较高的相似度，簇间具有较低的相似度，其相似度计算是根据一个簇中对象的平均值来进行。首先随机选择 n 个对象，每个对象代表了一个簇的初始平均值或中心。对于每个剩余的对象，根据它们与各个簇中心的距离赋给最近的簇，然后重新计算每个簇的平均值。不断重复这个过程，直到标准测度函数收敛。关于 K-MEANS 算法的详细介绍见 5.4.1 节。

（2）K-MEDOIDS：该方法首先为每个簇随意选择一个代表对象，对于每个剩余的对象，根据它们与代表对象的距离赋给最近的簇，然后反复地用非代表对象来替代代表对象，以改进聚类的质量。聚类结果的质量用一个代价函数来评估，该函数使用对象与其参照对象之间的平均相异度来度量。

3）基于密度的方法

这种方法是根据对象周围的密度来完成对象的聚类。典型的方法有：

（1）DBSCAN（Density-Based Spatial Clustering of Application with Noise）：该方法将一个聚类定义为一组“密度连接”的点的集合，将具有足够高密度的区域划分为簇，并能够从含有噪声的空间数据库中发现任意形态的聚类。

（2）OPTICS（Ordering Points To Identify the Clustering Structure）：该方法没有显式地产生一个数据集合簇，而是为自动和交互的聚类分析计算出一个簇顺序，这个顺序代表了数据的基于密度的聚类结构，它包含的信息等同于从一个宽泛的参数设置范围所获得的基于密度的聚类。OPTICS 方法解决了 DBSCAN 方法中存在的参数设置依靠经验、难以确定的难题。

4）基于网格的方法

这种方法首先将对象空间划分为有限个单元，以构成网格结构，然后利用网格结构完成聚类。典型的方法有：

（1）STING（STatistical INformation Grid）：该方法将空间区域划分为矩形单元，针对不同级别的分辨率，通常存在多个级别的矩形单元，这些单元构成一个层次结构，高层的每个单元被划分为多个低一层的单元。每个网格单元属性的平均值、最大值和最小值等统计信息被预先计算和存储，利用网格单元保存的统计信息进行聚类。

（2）CLIQUE（CLustering In QUEst）：该方法将基于网格的方法和基于密度的方法结合起来，给定一个多维数据点的大集合，数据点在数据空间中通常不是均衡分布的，通过区分数据空间中稀疏的和拥挤的区域或单元来发现数据集合的全局分布模式。如果一个单元中包含的数据点超过某个输入模型参数，则该单元是密集的。其工作过程分成两步，第一步将 n 维划分为互不相交的长方形单元，对于每一维都要识别其中的密集单元。第二步为每个簇生成最小化的描述，对每个簇，确定覆盖相连的密集单元的最大区域，然后确定最小的覆盖。

（3）Wave-Cluster：该方法是一种多分辨率的聚类算法，采用小波变换聚类，首先通过在数据空间上强加一个多维网格结构来汇总数据，每个网格单元汇总了一组映射到该单元中的点的信息，在多分辨率小波变换和聚类分析中需要使用这种汇总信息。然后采用一种小波变换来改变原来的特征空间，在变换后的空间中找到密集区域。

5）基于模型的方法

这种方法首先假设每个聚类的模型，然后挖掘出适合相应模型的数据。典型的方法有：

（1）COBWEB：该方法是一个常用且简单的增量式概念聚类方法，它的输入对象是用分类属性-值对来描述的，采用分类树的形式来创建一个层次聚类。

（2）CLASSIT：该方法是 COBWEB 方法的扩展，可以对连续取值属性进行增量式聚类。它为每个节点中的每个属性保存相应的连续正态分布（平均值与标准方差），并利用一个改进的分类描述方法，对连续属性求积分，而不像 COBWEB 那样在离散属性上求和。CLASSIT 和 COBWEB 方法存在同样的问题，它们都不适合对大规模数据集进行聚类分析。

5.3.5 文本分类

话题跟踪的任务是判断某个报道是描述了一个新话题还是对某个旧话题的进一步跟踪报道。话题跟踪可以看作一种特殊的文本分类过程，与传统的文本分类过程相比，话题跟踪中的文本分类是面向话题而不是面向概念更宽泛的主题，判断的依据更具体、粒度更细，处理的对象是动态的、随时间变化的报道流，而不是静态的文本集合。因此，在话题检测和跟踪中，不遗漏信息更为重要。

与文本聚类不同，文本分类是一种有监督的学习过程，需要事先给定一个分类体系和一个标注好类别的文本集合，利用这些资源来构造一个分类器，将待分类文本归入不同的、预先定义的类别中，可以把这种分类过程称为归类。

文本分类过程可以分为手工分类和自动分类，手工分类的著名实例是 Yahoo 网站的网页分类方法，它由专家定义分类体系，然后由人工进行网页分类。这种方法需要投入大量的人力，在现实中已经很少采用了。自动分类方法大致可以分为两类：知识工程方法和机器学习方法。知识工程方法是指由专家为每个类别定义一些规则，这些规则代表了这个类别的特

征，自动把符合规则的文档划分到相应的类别中。机器学习方法与知识工程方法相比，能够达到相似的精确度，同时还减少了大量的人工参与，成为文本分类的主流方法。下面介绍几种基于机器学习的文本分类方法。

1．文本分类的步骤

典型的文本分类过程可以分为以下 3 个步骤：

（1）文本表示。把文本表示成分类器能够处理的形式，最常用的方法是向量空间模型。

（2）分类器构建。选择或设计分类器构建方法，应当根据所要解决问题的特点来选择一个分类器。在选定方法之后，在训练集上为每个类别构建分类器，然后把分类器应用于测试集上，得到分类结果。

（3）效果评估。当分类算法在测试集（而不是训练集）上完成分类过程后，需要对算法的分类效果进行评估，常用的评价指标有准确率、召回率、漏报率和误报率等。

除了最简单的训练集－测试集评估方法外，还有一种 k-fold cross validation 方法，把所有标记的数据集划分成 k 个子集，对于每个子集，把该子集当作训练集，把其余子集作为测试集，这样执行 k 次，取各次评估结果的平均值作为最后的评估结果。

2．文本分类方法

在文本分类中使用的学习算法有多种，如 Rocchio 算法、KNN、决策树、简单贝叶斯、神经网络、最大熵、SVM 等。其中，比较常用的算法有 Rocchio 算法、KNN、决策树、SVM 等。

下面介绍几种常用的分类算法。

1）Rocchio 算法

该算法的基本思想是使用训练集为每个类构造一个原型向量，构造方法如下：

（1）给定一个类，训练集中所有属于这个类的文档对应向量的分量用正数表示，所有不属于这个类的文档对应向量的分量用负数表示，然后把所有的向量加起来，就是这个类的原型向量。

（2）采用余弦系数法计算训练集中所有文档和原型向量的相似度，然后按照一定的规则从中挑选某个相似度作为阈值。

（3）给定一个文档，如果这个文档与原型向量的相似度比阈值大，则该文档属于这个类，否则该文档就不属于这个类。

Rocchio 算法的基本思想可以解释为，对于一个词汇集和一个分类，词汇集中的某些词一旦出现，属于这个分类的可能性就会增加；而另一些词一旦出现，属于这个分类的可能性就会降低，统计这些正面和负面的影响因素，最后由文档分离出的词向量可以得到一个对于每个类的评分，分数越高属于该类的可能性就越大。

Rocchio 算法特别适合于某种非此即彼的分类，例如有两个类别：*A*、~*A*，任意给定一个文档，判断属于分类 *A* 还是分类~*A*，可以将 *A* 的特征项均赋予正值，将~*A* 的特征项都赋予负值，那么给定一个合理阈值，就很容易做出这种类型的分类。

Rocchio 算法的突出优点是容易实现，训练和分类的计算过程比较简单，通常被用来实现衡量分类系统性能的基准系统，而实用的分类系统很少采用这种算法来解决具体的分类问题。

2）KNN 算法

KNN（K-Nearest Neighbor）算法是一种比较成熟且简单的机器学习方法，该方法的基本思路是，如果一个样本在特征空间中的 *k* 个最相似（即特征空间中最邻近）样本的大多数都属于某一个类别，则该样本也属于这个类别。在 KNN 算法中，所选择的邻居都是已经正确分类的对象，在类别决策上只依据最邻近的一个或几个样本的类别来决定待分样本所属的类别。KNN 算法虽然从原理上也依赖于极限定理，但在类别决策时，只与极少量的相邻样本有关。由于 KNN 方法主要通过周围有限的邻近样本来确定所属类别，而不是依靠对类域的判别，因此对于类域交叉或重叠较多的待分样本集来说，KNN 方法是比较合适的方法。

KNN 算法用于分类时存在的主要问题如下：

当样本不平衡时，如一个类的样本容量很大，而其他类的样本容量很小时，有可能导致当输入一个新样本时，该样本的 *k* 个邻居中大容量类的样本占多数。这个问题可以采用加权的方法来改进，如样本距离小的邻居权值大。

该算法的计算量较大，因为对每一个待分类的文本都要计算它到所有已知样本的距离，这样才能求得它的 *k* 个最近邻居。这个问题的解决方法是事先对已知样本进行剪辑，去除对分类作用不大的样本。

该算法比较适合于样本容量较大的类域的自动分类，对于样本容量较小的类域容易产生误分。

KNN 算法不仅可以用于分类，还可以用于回归。例如，通过找出一个样本的 *k* 个最近邻居，将这些邻居的属性的平均值赋予该样本，就可以得到该样本的属性。

关于 KNN 算法的详细介绍见 5.5.1 节。

3）决策树算法

决策树算法是一种有序的贪心算法，其每一步都试图最大程度减小系统熵。决策树的构造方法是选择具有最大信息增益的特征作为根节点，根据该特征的值将训练样本分为不同的子集，对不同的子集重复进行，直到构造成一棵决策树。决策树生成后，便可以得到分类规则，待处理数据就可以根据该规则进行分类。常用的决策树算法有如 ID3、C4.5、CART 等。关于决策树算法的详细介绍见 3.3.1 节。

决策树算法用于话题跟踪时存在的主要问题是只能输出“是”和“否”的判断结果，而不能输出一个动态变化的可信度值。

事实上，每种分类算法都有各自的长处和局限性，它们经常可以互为补充。实际应用和算法实验表明，在文本分类中，KNN 方法或多种方法的组合具有较好的性能。

5.4　话题检测算法

话题检测的目的是按照话题对文档进行聚类，从一组新闻报道中发现新话题，它是在事先没有分类体系和训练语料的情况下对报道进行聚类分析，因此它是一个无监督学习的聚类问题。聚类算法有很多种，本节主要介绍三种聚类算法：K-MEANS 算法、FCM 算法和蚁群聚类算法，并通过实验来验证这些算法的性能。

5.4.1　K-MEANS 算法

K-MEANS 聚类法是一种简便实用的无监督学习算法，能够用于对已知类别的数据聚类和分类。理论和实验都证明，K-MEANS 聚类法是一种“理论上可靠、应用上高效”的聚类方法。

1．K-MEANS 算法的基本思想

该算法的基本思想是，把 n 个对象划分成 k 个类，其中聚类数量 k 是输入参数，通过不断地迭代来聚类，当算法收敛于一个结束条件时，终止迭代过程，输出一个聚类结果。

算法的工作过程如下：首先从 n 个数据对象中任意选择 k 个对象作为初始聚类中心，其他对象则根据它们与这些聚类中心的相似度（距离）分别分配给最相似的聚类，即聚类中心所代表的聚类。然后，计算每个获得新聚类的聚类中心，即该聚类中所有对象的均值，并不断重复这一过程，直到标准测度函数收敛为止，一般采用均方差作为标准测度函数。k 个聚类具有以下特点：各聚类本身尽可能紧凑，各聚类之间尽可能分开。

2．K-MEANS 算法的处理过程

K-MEANS 算法步骤描述如下。

算法 5-1　K-MEANS 算法

算法：K-MEANS，算法基于簇中对象的平均值。

输入：簇的数量 k 和包含 n 个对象的数据库。

输出：平方误差总和最小条件下的 k 个簇。

方法：

（1）任意选择 k 个对象作为初始的簇中心；

（2）将所有对象划分到相应的簇中；

（3）计算每个簇中对象的平均值，将所有对象重新赋给最类似的簇；

（4）重复（3），直到不再发生变化，收敛为止；

（5）算法结束。

K-MEANS 算法的处理过程实质上是一个逐步求精的迭代过程，如图 5-2 所示。首先任意选择 k 个对象作为初始的簇中心，按照种子间中垂线原理，将空间对象分配给所属的种子形成初始聚类，再对初始的每个簇求出其中心（均值），然后以此作为新的种子，重新聚类，直到所有的新种子不再更新为止。

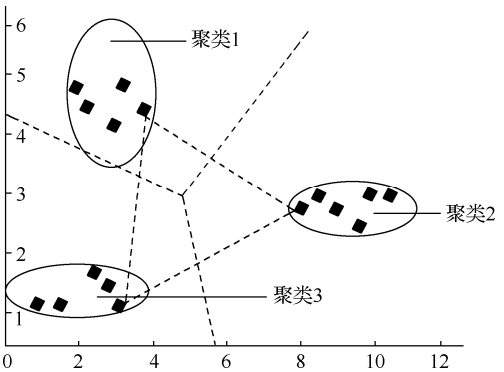


图 5-2　K-均值聚类过程

3．K-MEANS 算法的优化

在 K-MEANS 算法的处理过程中，簇的数量 k 是由算法的输入来提供的。那么对于一个实际的应用来说，多大的 k 值才是合适的，用户并不能提供一个准确的数值，仅能根据以往的经验或直觉提供一个大致的范围。在这种情况下，并不能直接运用 K-MEANS 算法来求解问题，应当先对 k 值进行优化处理，从而确定最优的 k 值。下面使用距离代价函数对 k 值进行优化处理。

定义 5-1　假设 n 个空间对象被聚类为 T_r 个簇，定义类间距离为所有分中心（每个簇的均值）到全域中心（所有空间对象均值）的距离之和，即：

$$L=\sum_{i=1}^{T_r}\left|\overline{u}_i-\overline{u}\right| \tag{5-11}$$

式中，L 为类间距离，$\overline{u}$ 为全部文档向量的均值，$\overline{u}_i$ 为簇 C_i 所含文档向量的均值，T_r 为聚类的个数，且 $T_r\in T$，k 为聚类个数的范围。

定义 5-2　假设 n 个空间对象被聚类为 T_r 个簇，定义类内距离为所有聚类内部距离的总和，其中每个聚类的内部距离为该聚类所有空间对象到其中心的距离之和，即：

$$D=\sum_{i=1}^{T_r}\sum_{u\in C_i}\left|\overline{u}-\overline{u}_i\right| \tag{5-12}$$

式中，D 为类内距离，v 为任一文档向量，T_r、$\overline{u}_i$、C_i 的含义与式（5-11）相同。

定义 5-3　定义距离代价函数为类间距离与类内距离的加权和，当该函数达到最小值时，表明以距离代价为依据的聚类结果为最优，即：

$$S=\omega_1 L+\omega_2 D \tag{5-13}$$

式中，S 为距离代价，L 和 D 分别为类间距离和类内距离，ω_1 和 ω_2 为加权指数，且满足 $\omega_1+\omega_2=1$（$\omega_1\leqslant\omega_2$）。

选择距离代价函数作为空间聚类有效性检验函数时，确定了距离代价最小准则，即当距离代价函数达到最小值时，空间聚类结果为最优。该准则源于空间聚类的一般原则，类别的划分应使得同一类（簇）的内部相似度最大、差异度最小，而不同类（簇）相似度最小、差异度最大。加权指数 ω_1 和 ω_2 的引入是为了区别类间距离 L 和类内距离 D 对于距离代价 S 的贡献度，经过多次实验，得出 $\omega_1\leqslant\omega_2$，即类内距离 D 对于距离代价 S 的贡献要大于类间距离 L，且获得如下经验值：$\omega_1=0.235$，$\omega_2=0.765$。

经过优化处理的 K-MEANS 算法称为优化 K-MEANS 算法，算法步骤描述如下。

算法 5-2　优化 K-MEANS 算法

算法：在 K-MEANS 算法基础上，通过距离代价函数优化 k 值。

输入：由用户给定簇的数目 $T_r(T_r\in T)$和包含 n 个对象的数据文件。

输出：距离代价函数最小条件下的 T_r 个簇。

方法：

（1）用 K-均值算法实现 T_r 个簇的聚类；

（2）根据距离代价函数分别计算不同聚类数目下的 S 值；

（3）搜寻距离代价函数最小的 S 值，并记录相应的聚类数目 k 值；

（4）算法结束。

经过优化 K-MEANS 算法处理后，得到了优化的 k 值，尽管并非是最优解，但提高了 k 值选择的准确性，从而提高了算法的准确度和效率。

5.4.2　模糊聚类方法

模糊聚类方法是建立在模糊数学基础上的聚类方法，与传统的聚类方法不同，模糊聚类方法打破了“非此即彼”的思想，而是采用了“亦此亦彼”的思想，更适合人类的思维模式。

从实现方法上，模糊聚类方法大致可以分为如下 4 类：

（1）谱系聚类法：该方法又称为 HCM（Hierarchical Clustering Method）方法，将样本集按照某种距离准则逐步分类，类别由多到少，直到满足一定的分类要求为止。在该算法中，常用的距离准则有最短距离法、最长距离法、中间距离法、重心法、类平均距离法等。

（2）基于等价关系的模糊聚类算法：该方法是通过建立模糊等价关系对样本集中的样本进行分类，在建立等价关系的过程中，相似度计算非常重要。目前，人们给出了一些常用的样本相似度计算方法。如余弦系数法、数量积法、相关系数法、指数相似系数法、最大最小法、绝对值减数法等。

（3）图论聚类法：该方法是通过构造支撑树来聚类的。

（4）基于目标函数的模糊聚类算法：该方法把聚类分析归结成一个带约束的非线性规划问题，通过优化求解获得数据集的最优模糊划分和聚类。该方法设计简单，解决问题范围广，还可以转化为优化问题并借助于经典数学的非线性规划理论来求解，并易于计算机实现。

1．FCM 算法

FCM（Fuzzy C-Means）算法是一种基于划分的聚类算法，其思想是使得被划分到同一簇的对象之间的相似度最大，而不同簇之间的相似度最小。FCM 算法是普通 C 均值算法的改进，普通 C 均值算法对于数据的划分是硬性的，而 FCM 算法则是一种柔性的模糊划分。

FCM 算法把 n 个向量 x_i（i = 1，2，…，n）分为 c 个模糊组，并求出每组的聚类中心，使得非相似度指标的价值函数达到最小。FCM 算法采用模糊划分，对于每个给定数据点，用值在 0，1 间的隶属度来确定其属于各个组的程度。与引入模糊划分相适应，隶属矩阵 U 允许有取值在 0，1 间的元素，通过归一化，一个数据集的隶属度之和总等于 1，即：

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \cdots, n \tag{5-14}$$

那么，FCM 的价值函数（或目标函数）为：

$$J(U, c_1, \cdots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{5-15}$$

式中，u_{ij} 介于 0，1 间，c_i 为模糊组 I 的聚类中心，d_{ij} = $\|c_i - x_j\|$ 为第 i 个聚类中心与第 j 个数据点间的欧几里得度量，且 $m \in [1, \infty)$ 是一个加权指数。

构造如下新的价值函数，可求得使式（5-15）达到最小值的必要条件：

$$\bar{J}(U,c_1,\cdots,c_c,\lambda_1,\cdots,\lambda_n)=\sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^{m}d_{ij}^{2}+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) \tag{5-16}$$

式中，$\lambda_j(j=1,\cdots,n)$是式（5-14）的 n 个约束式的拉格朗日乘子。

对所有输入参量求导，使式（5-15）达到最小值的必要条件为：

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{5-17}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{5-18}$$

由上述的两个必要条件可以看出，FCM 算法是一个简单的迭代过程。在批处理方式运行时，FCM 算法采用下列步骤来确定聚类中心 c_i 和隶属矩阵 U：

（1）使用值在 0、1 间的随机数初始化隶属矩阵 U，使其满足式（5-14）中的约束条件。

（2）用式（5-17）计算 c 个聚类中心 c_i，$i=1$，…，c。

（3）用式（5-15）计算价值函数，如果它小于某个确定的阈值，或者相对上次价值函数值的改变量小于某个阈值，则算法停止。

（4）用式（5-18）计算新的 U 矩阵。返回步骤（2）。

上述算法也可以先初始化聚类中心，然后再执行迭代过程。由于不能确保 FCM 算法收敛于一个最优解，算法的性能依赖于初始聚类中心。因此，要么用另外的快速算法来确定初始聚类中心，要么每次用不同的初始聚类中心启动该算法，需要多次运行 FCM 算法。

2．FCM 算法参数

FCM 算法有两个非常重要的参数：聚类数量 C 和加权指数 m。通常，C 要远小于聚类样本的总数，同时还要保证 $C>1$。算法的输出是 C 个聚类中心点向量和一个 $C\times N$ 的模糊划分矩阵，这个矩阵表示每个样本点属于每个类的隶属度，根据这个矩阵并按照模糊集合中的最大隶属原则就能确定每个样本点归为哪个类，聚类中心点表示每个类的平均特征，可以认为是这个类的代表点。

加权指数 m 是控制算法柔性的参数，如果 m 取值不合适，不仅会影响 FCM 算法的收

敛性，而且会影响模糊聚类的分类性能。如果 m 取值过大，则聚类效果比较差；如果 m 取值过小，则算法将退化为 HCM 聚类算法。研究表明，m 的最佳选取区间为[1.5，2.5]，在没有特殊要求时，可取中间值 $m=2$。

5.4.3 蚁群聚类算法

蚁群聚类算法是蚁群算法在文本聚类中的应用，它利用个体与个体以及个体与环境的交互作用实现自组织聚类。由于蚁群中的每个蚂蚁都是独立进行聚类活动的，彼此互不影响，增加了聚类过程的并行性，提高了效率。由于蚂蚁之间彼此互不依赖，一个蚂蚁的失效不会影响全局，增强了文本聚类的健壮性。蚁群聚类算法可以把对多维的文本向量聚类结果显示在两维的坐标系中，使得结果更加直观，增加了文本聚类的可视性。因此，蚁群聚类算法不仅实现了自组织聚类，还具有并行性、健壮性和可视化等方面的优点。

蚁群聚类算法的基本思想如下：

（1）把一定数量的文档分布在 $m \times m$ 的网格区域内，其中 m 是区域的宽度，要求每个网格内最多放置一个文档，m 的大小以及网格的数量随文档的数量而定；

（2）把一定数量的蚂蚁也分布在网格中，每只蚂蚁随机选择一个文档，根据该文档在局部区域的相似度而得到概率，决定蚂蚁是否“拾起”、“移动”或者“放下”该文档；

（3）经过有限次的迭代，平面区域内的文档按其相似度而聚集。

当蚂蚁发现一个文档后，采用下式来计算该文档与邻域内文档的群体相似度。

$$F(d_i)=\frac{\sum_{d_j\in \mathrm{round}(d_i)}\frac{1-\mathrm{sim}(d_i\times d_j)}{a}}{\pi r^2} \tag{5-19}$$

式中，round 是邻域的大小，它是以文档 d_i 为中心，以 r 为半径的一个圆形区域，r 也称为蚂蚁的观察半径，区域内的文档都在计算范围之内。a 是群体相似系数，是一个 1～10 的整数，它会影响最终的聚类个数并决定了算法的收敛速度。a 越大，对象间的相似度 F 越大，这样就会使不太相同的对象归为一类，聚类数量越少，收敛速度也就越快。反之，a 越小，对象之间的相似度越小，在极端的情况下，可能将一个大类分成了若干个小类，同时，随着聚类数量的增多，收敛速度将变慢。$\mathrm{sim}(d_i \times d_j)$是空间内两个物体的距离，在文本聚类中，它是指两个文本之间的相似度，通常采用欧几里得距离来度量，计算公式如下。

$$\mathrm{sim}(d_i,d_j)=\sqrt{\sum_{k=1}^{m}(w_{ik}-w_{jk})^2} \tag{5-20}$$

式中，w_{ik} 表示文档 d_i 中第 k 个关键词的权重，m 表示文档中关键词的个数。

$F(d_i)$是一个概率转换函数，可以将文档的群体相似度转化为“拾起”概率 P_p 或者“放下”概率 P_d，转换原则为：假设蚂蚁此时没有背负物体，如果此时蚂蚁选择的物体与周围的物体相似度越小，被拾起的可能性越大；选择的物体与周围的物体相似度越大，被拾起的可能性越小。如果此时蚂蚁背负物体，则蚂蚁会试图放下物体，如果此时蚂蚁背负的物体与周围的物体相似度越大，被放下的可能性越大；背负的物体与周围的物体相似度越小，被放下的可能性越小。拾起概率 P_p 和放下概率 P_d 分别是 F 的线性函数，计算方法如下：

$$P_p=\begin{cases}1 & F(d_i)\leqslant 0\\ 1-k\times F(d_i) & 0<F(d_i)\leqslant 1/k\\ 0 & F(d_i)>1/k\end{cases} \tag{5-21}$$

$$P_d=\begin{cases}0 & F(d_i)\leqslant 0\\ k\times F(d_i) & 0<F(d_i)\leqslant 1/k\\ 1 & F(d_i)>1/k\end{cases} \tag{5-22}$$

上式反映了 F 与拾起概率 P_p 或放下概率 P_d 是线性关系，k 是直线的斜率。

蚁群聚类算法步骤描述如下。

算法 5-3　蚁群聚类算法

（1）初始化文档，把 N 个文档随机分布在 $m\times m$ 的网格内，要求 $m\times m>N$；

（2）初始化蚂蚁，选 K 只蚂蚁放入网格区域内，为蚂蚁分配文档；

（3）在蚂蚁的观察半径内，利用式（5-19）计算 F；

（4）如果蚂蚁没有背负文档，利用式（5-21）计算拾起概率 P_p。如果 $P_p\geqslant P$，则拾起文档，并为蚂蚁随机分配一个新文档；如果 $P_p<P$，则不拾起，并为蚂蚁随机分配另一个新文档；

（5）如果蚂蚁背负文档，利用式（5-22）计算放下概率 P_d。如果 $P_d\geqslant P$，则放下文档，并为蚂蚁随机分配一个文档；如果 $P_d<p$，则不放下，并为蚂蚁随机分配另一个新文档；

（6）到达循环次数，算法结束，显示聚类结果。

蚁群聚类算法流程图如图 5-3 所示。

5.4.4　算法验证

下面通过实验数据对 K-MEANS 算法、FCM 算法和蚁群聚类算法的性能进行测试和验证。

1．实验数据集

实验数据是通过网络爬虫工具从新浪网站（www.sina.com.cn）和搜狐网站（www.sohu.com）上下载了 2 287 篇新闻网页，包含了 14 个话题，发生的时间围是从 2008 年 1 月到 2009 年 2 月期间，涵盖了经济、政治、生活等多个方面，其事件分布情况如表 5-1 所示。

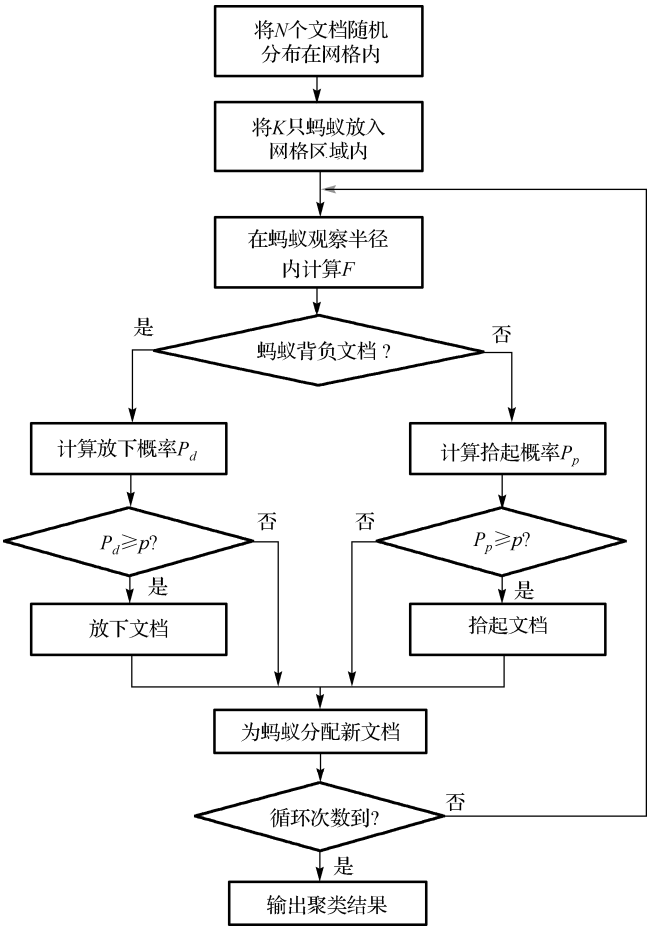


图 5-3　蚁群聚类算法流程图

表 5-1　实验数据集的事件分布情况

序　号	话题名称	新闻报道总数	训练集大小	测试集大小
1	中国雪灾	200	80	120
2	四川大地震	200	80	120
3	北京奥运会	236	80	156
4	美国第一位黑人总统的诞生——奥巴马	106	80	26
5	2008 金融危机	178	80	98
6	雷曼兄弟公司破产	100	80	20
7	冰岛破产	100	80	20
8	俄罗斯中断对乌克兰天然气供应	134	80	54
9	世界经济论坛 2009 年年会开幕	143	80	63

续表

序　号	话题名称	新闻报道总数	训练集大小	测试集大小
10	甲型 H1N1 流感疫情在多国蔓延	200	80	120
11	朝鲜进行核试验	146	80	66
12	通用汽车公司申请破产保护	127	80	47
13	“世纪日食”观测热潮	131	80	51
14	韩国首枚运载火箭发射出现异常	100	80	20

新闻报道所属话题信息全部由手工标注。其中，在每个话题下选择前 80 篇作为训练集的语料，剩下的作为测试集的语料。由于本测试的语料与 TDT 项目测评的语料不同，其测试结果不能与 TDT 项目的测评结果相比较。

2．K-MEANS 算法验证

在 K-MEANS 算法实验中，设置隐藏话题的数量 k 为 14，表 5-2 给出了 K-MEANS 算法对 14 个话题的检测准确率、召回率、漏报率、误报率和$(C_{\mathrm{Det}})_{\mathrm{Norm}}$值。其中，归一化识别代价$(C_{\mathrm{Det}})_{\mathrm{Norm}}$值是根据漏报率和误报率计算得到的，见式（5-1）。

表 5-2　K-MEANS 算法对 14 个话题的检测性能

序　号	话题名称	准确率	召回率	漏报率	误报率	$(C_{\mathrm{Det}})_{\mathrm{Norm}}$
1	中国雪灾	0.685	0.784	0.216	0.035	0.219
2	四川大地震	0.675	0.694	0.306	0.031	0.312
3	北京奥运会	0.798	0.787	0.213	0.01	0.219
4	美国第一位黑人总统的诞生——奥巴马	0.792	0.758	0.242	0.011	0.248
5	2008 金融危机	0.705	0.792	0.208	0.038	0.213
6	雷曼兄弟公司破产	0.712	0.656	0.344	0.039	0.349
7	冰岛破产	0.698	0.674	0.328	0.037	0.328
8	俄罗斯中断对乌克兰天然气供应	0.798	0.743	0.257	0.015	0.261
9	世界经济论坛 2009 年年会开幕	0.772	0.761	0.239	0.028	0.242
10	甲型 H1N1 流感疫情在多国蔓延	0.768	0.685	0.315	0.013	0.32
11	朝鲜进行核试验	0.725	0.751	0.249	0.026	0.253
12	通用汽车公司申请破产保护	0.7072	0.693	0.307	0.038	0.314
13	“世纪日食”观测热潮	0.792	0.766	0.235	0.011	0.238
14	韩国首枚运载火箭发射出现异常	0.735	0.709	0.295	0.023	0.297

下面从话题检测的准确率、召回率、漏报率、误报率等方面对实验结果进行分析。

从表 5-2 可以看出，14 个话题的检测准确率为 67%～80%，其中有 4 个话题的检测准确率比较高，大于 79%。“中国雪灾”、“四川大地震”、“冰岛破产”等话题检测的准确率比较低（小于 70%），其主要原因有如下几个方面：

（1）这些话题本身的特殊性，它们都不是独立性很强的话题，而是包含了若干个子话

题。例如，在“四川大地震”这个话题中，包含了对四川地震的实时报道、对地震原因的讲解、对伤员的救护、对幸存人员的营救、对抗震英雄事迹的报道以及灾后的重建等若干个子话题。因此，在处理这个话题时，话题的概念边缘被扩大了，认为这个话题表述的内容范围比较广，被误分到这个话题下的概率就增加了，从而造成准确率的下降。这也说明了在多层次话题检测时，还需要改进其准确率。

（2）由于类别相同或相似的话题之间可能存在一定的交叉性关键词，也影响到话题检测准确率。例如，“四川大地震”和“中国雪灾”都属于灾害类新闻话题，都可能出现诸如“灾后重建”、“救灾款项”等之类的关键词，给区分这类话题带来一定的困难。另外，“2008 金融危机”、“通用汽车公司申请破产保护”等话题，同样存在相同关键词的相互干扰问题，影响到其检测准确率。

（3）K-MEANS 算法对检测时间跨度较短的话题具有较好的性能，而检测像“四川大地震”这类时间跨度比较长的话题，它的性能比较差。

14 个话题检测的召回率为 65%～80%，其中，“中国雪灾”、“北京奥运会”、“2008 金融危机”等话题检测的召回率到达了 78%以上，而“雷曼兄弟破产”、“冰岛破产”话题检测的召回率都低于 68%，“雷曼兄弟破产”、“冰岛破产”、“2008 金融危机”等话题都与金融危机相关，话题之间形成了相互干扰，有一部分属于“雷曼兄弟破产”、“冰岛破产”主题的新闻报道被误判给“2008 金融危机”主题了。

漏报率与召回率是互补的，即：漏报率 = 1–召回率。而误报率的高低则取决于算法对多层次话题的识别能力，在 14 个话题中，有些话题检测的误报率比较高，超过了 3%。

总之，K-MEANS 算法是一种属于“非此即彼”的硬性划分，即一个报道不是属于这个话题，就是属于那个话题，当一个报道同时与两个话题相似时，K-MEANS 算法就会出现误分问题，降低了算法的检测性能。

3．FCM 算法验证

将 FCM 算法应用于实验数据集，得到的 14 个话题的检测准确率、召回率、漏报率、误报率和$(C_{Det})_{Norm}$值，如表 5-3 所示。

表 5-3　FCM 算法对 14 个话题的检测性能

序　号	话 题 名 称	准确率	召回率	漏报率	误报率	$(C_{Det})_{Norm}$
1	中国雪灾	0.725	0.764	0.236	0.022	0.219
2	四川大地震	0.734	0.748	0.252	0.024	0.312
3	北京奥运会	0.797	0.782	0.218	0.015	0.219
4	美国第一位黑人总统的诞生——奥巴马	0.79	0.75	0.25	0.019	0.258
5	2008 金融危机	0.739	0.787	0.213	0.029	0.219
6	雷曼兄弟公司破产	0.753	0.694	0.306	0.027	0.314
7	冰岛破产	0.732	0.708	0.292	0.03	0.302

续表

序号	话 题 名 称	准确率	召回率	漏报率	误报率	$(C_{Det})_{Norm}$
8	俄罗斯中断对乌克兰天然气供应	0.789	0.725	0.275	0.013	0.278
9	世界经济论坛 2009 年年会开幕	0.784	0.758	0.242	0.024	0.248
10	甲型 H1N1 流感疫情在多国蔓延	0.786	0.709	0.291	0.013	0.296
11	朝鲜进行核试验	0.738	0.743	0.257	0.021	0.262
12	通用汽车公司申请破产保护	0.737	0.705	0.295	0.028	0.306
13	“世纪日食”观测热潮	0.798	0.772	0.228	0.01	0.231
14	韩国首枚运载火箭发射出现异常	0.747	0.727	0.273	0.022	0.28

从表 5-3 可以看出，与 K-MEANS 算法相比，FCM 算法的检测准确率有了一定的提升，但提升的幅度并不大。对于交叉性话题（即话题之间存在着较多的交叉性关键词），如“四川大地震”、“中国雪灾”，FCM 算法的检测准确率有明显的提高；对于像“北京奥运会”等独立性较强的话题，FCM 算法的检测准确率并没有明显的提高，与 K-MEANS 算法持平，因为这种话题的模糊性比较小，就是一个“非此即彼”的关系。在召回率上，针对不同的话题，FCM 算法有升有降，总体上与 K-MEANS 算法差别不大。

FCM 算法是一种糊聚类方法，属于柔性划分算法，比较适合于处理不确定性对象的划分问题。对于独立性较强的话题，其模糊性比较小，就是一个“非此即彼”的关系，其算法性能与硬性划分算法没有太大的差别。另外，FCM 算法与硬性划分算法之间存在着联系，如果 FCM 算法中的加权指数 m 值取得过小，则算法性能就十分接近于硬性划分算法。因此，FCM 算法的 m 参数取值非常重要。

4．蚁群聚类算法验证

在蚁群聚类算法中，需要设置一些参数，如表 5-4 所示。

表 5-4　蚁群聚类算法参数的设置

参数名称	参数含义	参　数　值
蚂蚁数	决定程序的并发度	50
a	群体相似系数	3
拾起概率参数 k_1	其值决定蚂蚁是否拾起文档	0.4
放下概率参数 k_2	其值决定蚂蚁是否放下文档	0.15
观察半径 R	R 内的文档都在计算相似度范围之内	8

将蚁群聚类算法应用于实验数据集，得到的 14 个话题的检测准确率、召回率、漏报率、误报率和$(C_{Det})_{Norm}$值，如表 5-5 所示。

表 5-5　蚁群聚类算法对 14 个话题的检测性能

序　号	话题名称	准确率	召回率	漏报率	误报率	$(C_{Det})_{Norm}$
1	中国雪灾	0.437	0.815	0.185	0.064	0.226

续表

序　号	话题名称	准确率	召回率	漏报率	误报率	$(C_{Det})_{Norm}$
2	四川大地震	0.395	0.325	0.675	0.032	0.692
3	北京奥运会	0.693	0.793	0.207	0.013	0.213
4	美国第一位黑人总统的诞生——奥巴马	0.674	0.794	0.207	0.016	0.239
5	2008 金融危机	0.427	0.826	0.174	0.07	0.193
6	雷曼兄弟公司破产	0.383	0.328	0.672	0.043	0.694
7	冰岛破产	0.367	0.336	0.664	0.04	0.687
8	俄罗斯中断对乌克兰天然气供应	0.683	0.738	0.262	0.018	0.269
9	世界经济论坛 2009 年年会开幕	0.586	0.698	0.302	0.034	0.308
10	甲型 H1N1 流感疫情在多国蔓延	0.676	0.808	0.192	0.012	0.204
11	朝鲜进行核试验	0.538	0.809	0.191	0.02	0.2
12	通用汽车公司申请破产保护	0.327	0.312	0.688	0.044	0.702
13	“世纪日食”观测热潮	0.658	0.735	0.265	0.015	0.275
14	韩国首枚运载火箭发射出现异常	0.563	0.797	0.203	0.02	0.212

从表 5-7 可以看出，与 K-MEANS 算法和 FCM 算法相比，不论是交叉性的话题还是独立性较强的话题，蚁群聚类算法的检测准确率都出现了不同程度的下降，其主要原因是：

（1）每只蚂蚁是独立的个体，遇到的情况可能是不一样的，算法参数也不能按照统一的标准来设置，如比较概率，应当为每只蚂蚁设置相应的比较概率。

（2）聚类是一个过程，在此过程中，蚂蚁要和环境进行交流，要随着环境的变化做出自己的调整。因此蚂蚁的一些参数应当具有动态性，比如蚂蚁的观察半径等。

从表 5-5 可以看出，蚁群聚类算法的检测召回率起伏较大，有 4 个话题检测的召回率比较高，超过了80%，达到了较高的水平；另有 4 个话题检测的召回率比较低，低于34%。

蚁群聚类算法是一种群体决策的聚类算法，在聚类前可以不必预先设置聚类的数量，就能实现自组织聚类。由于没有聚类数量的限制，蚁群聚类算法可以根据局部相似度来实现自组织聚类，在处理某些情况可能会出现偏差，例如，对于“四川大地震”和“中国雪灾”这两个话题，蚁群聚类算法在处理时可能将两者合二为一，将有关“四川大地震”的报道归属于“中国雪灾”话题中，于是出现了“中国雪灾”话题的召回率高，而“四川地震”话题的召回率低的现象。

5．三种算法性能对比

根据表 5-2、表 5-3、表 5-5 中给出的三种算法漏报率、误报率和$(C_{Det})_{Norm}$ 值，分别计算出三种算法的平均漏报率、平均误报率和平均$(C_{Det})_{Norm}$ 值，比较三种算法的整体性能，如表 5-6 所示。

从表 5-6 可以看出，三种算法相比，FCM 算法的三项指标都比较低，而蚁群聚类算法

的三项指标都比较高，算法性能由高到低排序是 FCM 算法、K-MEANS 算法、蚁群聚类算法，因此，在三种算法中，选择 FCM 算法作为话题检测算法是比较适宜的。

表 5-6　三种算法的平均漏报率、误报率和$(C_{Det})_{Norm}$值

算法名称	平均漏报率	平均误报率	平均$(C_{Det})_{Norm}$值
K-MEANS 算法	0.268	0.025	0.272
FCM 算法	0.258	0.021	0.258
蚁群聚类算法	0.349	0.032	0.365

5.5　话题跟踪算法

话题跟踪是一个文本分类过程，与文本聚类不同，文本分类是一种有监督的学习过程，需要事先给定一个分类体系和一个标注好类别的文本集合，利用这些资源来构造一个分类器，将待分类文本归入不同的、预先定义的类别中。文本分类中使用的学习算法有多种，如 Rocchio 算法、KNN、决策树、简单贝叶斯、神经网络、最大熵、SVM 等，其中分类效果比较好的是 KNN 算法以及多种算法的组合。下面主要介绍 KNN 算法及其改进。

5.5.1　KNN 算法及改进

1．基本 KNN 算法

KNN 是一种基于机器学习的分类算法，其实质就是记忆，即把新的知识存储起来，供需要时使用，而不需要推理和计算。KNN 是性能比较好的文本分类算法之一，其他的较好的方法还有 SVM、决策树、神经网络等。

KNN 算法思想比较简单，对于给定的一个测试文档，在训练集中查找离它最近的 k 个邻居，并根据这些邻居的类别，给该文档的候选类评分，把邻居文档和测试文档的相似度作为邻居文档所在类的权重。如果这 k 个邻居中的部分文档属于同一个类别，则将该类中的每个邻居的权重之和作为该类和测试文档的相似度，通过对候选类评分的排序，然后给出一个阈值，就可以判定测试文档的类别。

话题跟踪实际上是二元分类问题，利用 KNN 算法进行话题跟踪的基本思路也比较简单，在给定一个新文本后，考虑在训练文本集中与该文本距离最近（最相似）的 k 个文本，根据这 k 个文本所属的正例、反例的相似度值大小来判定新文本是否属于该话题，具体的算法步骤如下：

（1）使用余弦系数法作为相似度度量方法，分别计算新文本 x 与训练集文本 d_i 的相似度值 $\cos(\vec{x},\vec{d}_i)$，选出与新文本最相似的 k 个文本。

（2）在这 k 个文本中，抽取属于正例样本集合 P_k 的所有文本，将这些文本与新文本 x

的相似度值 $\cos(\vec{x},\vec{u})(u \in P_k)$ 求和，作为新文本 x 与正例集合 P_k 的相似度值。同样，在最相近 k 个文本中抽取属于反例样本集合 Q_k 的所有文本，将这些文本与新文本 x 的相似度值 $\cos(\vec{x},\vec{u})(u \in Q_k)$ 求和，作为新文本 x 与反例集合 Q_k 的相似度值。

（3）比较新文本与正例、反例的相似度值，计算公式如下：

$$f_{\arg}(\vec{x} \mid k) = \sum_{\vec{u} \in p_k} \cos(\vec{x},\vec{u}) - \sum_{\vec{v} \in Q_k} \cos(\vec{x},\vec{v}) \tag{5-23}$$

（4）当 $f_{\arg}(\vec{x} \mid k) \geqslant 0$，判定新文本 x 属于该话题，当 $f_{\arg}(\vec{x} \mid k) < 0$，判定新文本 x 不属于该话题。

2．KNN 算法改进

在 KNN 算法中，k 是一个重要的参数，它的取值直接关系到算法的性能。由于训练语料比较稀疏，当算法中 k 值选取过大时，过多的反例就会影响正例的判断，增大话题跟踪的误报率和漏报率。

为了克服在话题跟踪中训练样本正例少而引起的偏差问题，对 KNN 算法进行如下的改进：首先在训练集中选出与新文本最相似的 k 个文本，在这 k 个文本中分别抽取 kp 个属于正例的文本和 kn 个属于反例的文本，且 $kp+kn=k$，构成正例集 U_{kp} 和反例集 V_{kn}，然后分别计算正例集、反例集中的文本与新文本的平均相似度，最后根据正、反例平均相似度值大小对新文本做出判断，计算公式如下：

$$f_{\arg}(\vec{x} \mid k_p, k_n) = \frac{\sum_{\vec{u} \in U_{kp}} \cos(\vec{x},\vec{u})}{|U_{kp}|} - \frac{\sum_{\vec{v} \in V_{kn}} \cos(\vec{x},\vec{v})}{|V_{kn}|} \tag{5-24}$$

KNN 算法在计算新文本 x 与训练集文本 d_i 的相似度时，相似度度量方法非常关键，将对算法性能产生重要影响。相似度度量方法有多种，除了上述的余弦系数法外。还有内积法、Dice 系数、Jaccard 系数、欧几里得度量等方法。

5.5.2 算法验证

下面通过实验数据对 KNN 算法性能进行测试和分析。

1．实验数据集

实验内容有两个：一是对 KNN 算法与改进 KNN 算法的性能进行对比，考察改进 KNN 算法对性能有多大的提升；二是对余弦系数、Dice 系数、Jaccard 系数、欧几里得度量 4 种相似度度量方法进行对比，考察不同的相似度度量方法对算法性能有多大的影响。

在实验中使用了如下的数据集：

（1）IRIS 数据集。该数据集是从不同植物的花特征提取出来的，该数据集共有 150 个

数据点，每个数据点有 4 个属性，它们分别是花萼的长度、花萼的宽度、花瓣的长度和花瓣的宽度。

（2）Wine 数据集：该数据集是从三种不同种类葡萄所酿制的葡萄酒中来提取的，共有 178 个样本，每个样本有 13 个属性，其中每种葡萄酒提取的数据个数分别为 34、71 和 48。

（3）WDBC 数据集：该数据集是有关乳腺癌病人的数据集，由 569 个样本组成，每个样本有 32 个属性，分别来自不同乳腺癌病人，其中第一个属性为病人的标识号（ID），第二个属性为癌症的性质（即良性还是恶性）。

这些数据集是开放的，可以从互联网下载得到。

2．两种 KNN 算法性能对比

在这里，KNN 算法是指原始的 KNN 算法，将改进后的 KNN 算法称为 $KNN_{(new)}$算法。将两种 KNN 算法应用于上述的数据集，分别选取 $N_t = 2$ 和 $N_t = 4$ 个训练正例进行测试，表 5-7 和表 5-8 分别为两种 KNN 算法 $N_t = 2$ 和 $N_t = 4$ 时的实验结果。

表 5-7　两种 KNN 算法在 $N_t = 2$ 时的实验结果

评价指标	KNN 算法	$KNN_{(new)}$算法
准确率	0.826	0.854
召回率	0.843	0.866
漏报率	0.157	0.134
误报率	0.019	0.019
$(C_{Det})_{Norm}$ 值	0.184	0.156

表 5-8　两种 KNN 算法在 $N_t = 4$ 时的实验结果

评价指标	KNN 算法	$KNN_{(new)}$算法
准确率	0.874	0.896
召回率	0.869	0.887
漏报率	0.131	0.113
误报率	0.018	0.015
$(C_{Det})_{Norm}$ 值	0.159	0.134

从表 5-7 可以看出，当 $N_t = 2$ 时，$KNN_{(new)}$算法的性能要优于 KNN 算法，这是因为 KNN 算法需要大量的训练语料，这样才能获得较好的算法性能。当训练语料过于稀疏时，如果 k 值选取过小，KNN 算法性能较低；如果 k 值选取过大，过多的训练反例将会影响算法的判断。$KNN_{(new)}$算法通过对正、反例相似度值求平均，降低了 k 值对 KNN 算法性能的影响，也减弱了训练反例对判断的影响，因此降低了漏报率和误报率，提高了准确率和召回率。

从表 5-8 可以看出，当 $N_t = 4$ 时，两种方法的实验结果均优于 $N_t = 2$ 时的实验结果，从中不难看出训练正例对 KNN 算法的影响。

实验结果表明，$KNN_{(new)}$算法比较好地解决了训练正例稀疏的问题，提高了算法的准确率，降低了漏报率和召回率。

3．相似度度量方法对算法性能影响

将 $KNN_{(new)}$算法分别与余弦系数、Dice 系数、Jaccard 系数、欧几里得度量 4 种相似度度量方法相结合，考察不同相似度度量方法对算法性能的影响，实验结果如表 5-9 所示。

表 5-9　不同的相似度度量方法对 $KNN_{(new)}$算法性能的影响

评价指标	余弦系数	Dice 系数	Jaccard 系数	欧几里得度量
准确率	0.854	0.814	0.822	0.874
召回率	0.866	0.826	0.825	0.871
漏报率	0.134	0.174	0.175	0.129
误报率	0.019	0.021	0.019	0.019
$(C_{Det})_{Norm}$ 值	0.156	0.197	0.196	0.150

从表 5-9 可以看出，不同的相似度度量方法确实对 $KNN_{(new)}$算法性能产生影响，其算法性能由高到低排序是欧几里得度量、余弦系数、Jaccard 系数、余弦系数，因此 $KNN_{(new)}$算法结合欧几里得度量或余弦系数度量方法，可以达到最佳的效果。

5.6　热点话题检测

下面以网络论坛突发性热点话题检测为例，介绍 TDT 技术在网络网络舆情分析中的应用。在网络论坛中，网民的观点是通过发帖或回帖进行传播的，在检测突发性热点话题时，需要考虑网络论坛结构、话题特征以及突发性热点话题所具备的高关注度和时间突发特性，采用噪声过滤、文本聚类等方法实现对突发性热点话题的检测与跟踪。

5.6.1　检测方法

面向网络论坛的突发性热点话题检测方法包括候选话题集构建、噪声过滤、热点话题检测、热点话题跟踪 4 个步骤。

（1）候选话题集构建：对于采集到的网络论坛数据，构建以主帖标题为索引的候选话题集；

（2）噪声过滤：对候选话题集进行去噪处理，过滤掉热度值较低的主帖和不具有时间突发性的主帖，这样可以过滤掉网络论坛中大部分不会演变成突发性热点话题的帖子；

（3）热点话题检测：通过分词工具提取主帖标题中所包含的主题词，并采用聚类方法对主帖进行合并，进而抽取出突发性热点话题；

（4）热点话题跟踪：针对突发性热点话题的时间序列，通过绘制其对应的回帖加速度变化曲线进行跟踪。

1．候选话题集构建

在网络论坛中，主帖的标题通常代表了用户讨论的主题。根据对 2011 年中国互联网舆情分析报告中列出的前 10 大网络热点事件，以影响力较大的网络论坛-天涯社区为对象，检索出该社区中对应 10 大网络热点事件的回复数最大的主帖，统计发现，能概括事件主题的主帖标题占 80%以上。因此，可以将网络论坛中的主帖标题作为索引来构建候选话题集。

2．噪声过滤

网络论坛的热点话题的形成分为内容驱动和形式驱动两种不同的方式。内容驱动是通过发表内容丰富的帖子来吸引大量的网民浏览和回复，从而形成热帖或热点话题；形式驱动是网络推手通过调动网络水军发表大量有关事件的帖子，将该事件强行推入公众视野，形成热点话题。

通过计算帖子热度值，给帖子热度评分，过滤掉热度值较低的帖子。帖子热度值计算公式如下：

$$\text{hotness}(x_i)=\alpha\frac{r(x_i)}{\text{avgr}(X)}+\beta\frac{b(x_i)}{\text{avgb}(X)}+\gamma\frac{r(x_i)/b(x_i)}{\max(X)} \tag{5-25}$$

式中，α、β 和 γ 均为加权值，计算得到 α = 0.1947、β = 0.0881、γ = 0.7172；$r(x_i)$为主帖 x_i 的回复数；$b(x_i)$为主帖 x_i 的点击数；avgr(X)为所有帖子 X 的平均回复数；avgb(X)为所有帖子 X 的平均点击数；max(X)为所有帖子 X 中最大回复数与点击数之比。

一般而言，用户必须点击某个帖子才能对它进行回复，因此帖子点击数一般要大于回复数，在式（5-25）中，以 $r(x_i)/b(x_i)$来描述用户围绕主帖 x_i 讨论的活跃程度，比值越大，表明该帖子相对应的话题关注度越高。

式（5-25）给出的帖子热度反映了帖子回复数和点击数的累积，但它不能体现帖子热度随时间的变化特性。根据网络论坛舆情演化理论，帖子的生命周期可分为突发、成长、衰退、消解 4 个阶段。处于衰退阶段的帖子即使其热度值很高，已不可能演变成具有时间突发性的热点话题。因此，采用回复加速度的概念来识别和量化帖子热度随时间的变化特性：

$$a(x_i)^t=\frac{r(x_i)^t-2r(x_i)^{t-\Delta t}+r(x_i)^{t-2\Delta t}}{(\Delta t)^2} \tag{5-26}$$

式中，$a(x_i)^t$ 为主帖 x_i 的回复加速度；$r(x_i)^t$ 为主帖 x_i 在时间 t 的回复数；Δt 为介于 $r(x_i)^t$ 之间的时间粒度。

同时，定义一个状态函数 $S(x_i)$：$R\rightarrow$ {acc, growth, dec, death}，以标识帖子生命周期中的不同阶段，该函数定义为：

$$S(x_i)^t=\begin{cases}\text{acc} & a(x_i)^t\in(\theta_1,+\infty)\\ \text{growth} & a(x_i)^t\in[\theta_2,\theta_1)\\ \text{dec} & a(x_i)^t\in(-\infty,\theta_2)\\ \text{death} & r(x_i)^t=0\end{cases}\tag{5-27}$$

式中，acc 表示帖子生命周期的突发期，growth 表示帖子生命周期的成长期，dec 表示帖子生命周期的衰退期，death 表示帖子生命周期的消解期，θ_1 和 θ_2 均为预先设定的阈值，$r(x_i)^t$ 为常量 0。

在突发期和衰退期，帖子的回复数或急剧增加（$a(x_i)>>0$）或强烈衰减（$a(x_i)<<0$）；在成长期，帖子在单位时间内的回复数基本不变；而在消解期，帖子回复数为 0。

通过对帖子的热度度量和热度随时间变化特性，不仅可以过滤掉热度值较低或已处于衰退、消解阶段的帖子，保留处于加速或成长阶段的帖子，还可以对回复加速度急剧增加的帖子给予足够的关注，而这些帖子往往容易发展成为突发性热点话题。

对去噪后保留下来的帖子建立主题集合 ST = {th_1, th_2, …, th_m}，其中 th_i 为主题词。

3．热点话题检测

在对主题进行聚类之前，需要对主题进行分词处理，以抽取出能够反映突发性热点话题特征的主题词。在中文分词中，采用中文分词系统进行分词。例如，对于“朝鲜今日宣布正计划进行第三次核试验”这一标题，通过分词后可以得到如下分词结果：朝鲜/*n* 今日/*t* 宣布/*v* 正/*d* 计划/*v* 进行/*v* 第三/*m* 次/*q* 核试验/*n*，抽取其中的名词和动词，删除重用词，可以得到集合{朝鲜，宣布，计划，进行，核试验}。

在实际操作中，对主题集合 ST 中的每个主题词 th_i 分别进行分词处理，并建立集合 th_i = {$term_j|1\leqslant j\leqslant n$}。同时，考虑到网络论坛中很多主题讨论的话题是现实中的新事件，如“房妹”，“房姐”等，在分词之前需要手动添加到中文分词系统的词库中。

经过分词处理后，抽取出突发性热点话题的主题词，基于主题词对热点话题进行检测。由于网络论坛中多个主题所讨论的话题可能是同一个热点话题，因此在同一话题的集合中可能包含有相同的主题词。例如，集合 th_1、th_2、th_3，分词后为 th_1 = {$term_1$, $term_2$, $term_3$, $term_4$}、th_2 = {$term_2$, $term_3$}、th_3 = {$term_1$, $term_2$, $term_3$}。在主题词进行聚类处理时，定义两个集合间的 Jaccard 系数作为它们的相似度度量，即：

$$\text{Sim}(th_i,th_j)=\frac{|th_i\cap th_j|}{|th_i\cup th_j|}\tag{5-28}$$

式中，$th_i\in ST$，$th_j\in ST$。

整个热点话题检测算法步骤描述如下：

算法 5-4　热点话题检测算法

输入：ST = {th_1, th_2, …, th_m} 和预设阈值 η。

输出：热点话题集合 H = {H (th_i)} 。

（1）初始化主题集合 ST，分为 m 个主题 th_1, th_2, …, th_m。

（2）对于 th_1, th_2, …, th_m，选取最大主题 max|th_i|，且有 sim (max|th_i|, th_j)≥ η，加入到热点话题集合 H (th_i) = {th_i, th_j} 中，否则转入（3）。

（3）ST = ST −H(th_i) ，迭代计算转入(2)，从 ST 中再次选择最大主题 max|th_i|，直到集合 ST 为空。

5.6.2　算法验证

下面通过实验数据对网络水军账号检测算法进行测试和验证。

1．实验数据集

实验数据来源于网易新闻论坛（http://bbs.news.163.com）2011 年 3 月 1 日-2011 年 5 月 1 日间的数据，该数据集的构成情况如表 5-10 所示。

表 5-10　实验数据集构成情况

类　　别	数　　量
主帖数	4 248
主帖 ID	1 562
主帖平均回复数	12
主帖平均点击数	1 811
最大回复数	550
最小点击数	1

事先对该数据集中的 4 248 个帖子进行人工标注，这些帖子共包含 2 716 个话题，并且有超过 2 000 个帖子是孤立的，而热点话题往往包含多个帖子。这说明在网络论坛中需要对热点话题进行归纳与总结，以方便用户及时了解发生的突发事件以及相关的事件。

2．算法有效性验证

根据对实验数据集中的 4 248 个主帖进行热度值计算，得到如图 5-4 所示的主帖热度时间分布图。

为了快速发现热帖，需要通过量化方法对该图的热帖区域进行划分。对于演变成热点话题的主帖，其对应的回帖数应当呈指数增长，即回帖数与热点话题的关系是指数关系而非线性关系。因此以主帖回复数与主帖平均回复数比值的对数，即 lg($r(x_i)$/avgr(X))作为横坐标，以 hotness>0.4318，lg($r(x_i)$/avgr(X))>1 作为热帖检测阈值，得到如图 5-5 所示的主帖热

度分布图。在图 5-5 中，通过阈值选择的热帖数目有 406 个，其中在纵横两条虚线构成的 4 个区域中，右上的区域即为检测出的热帖。

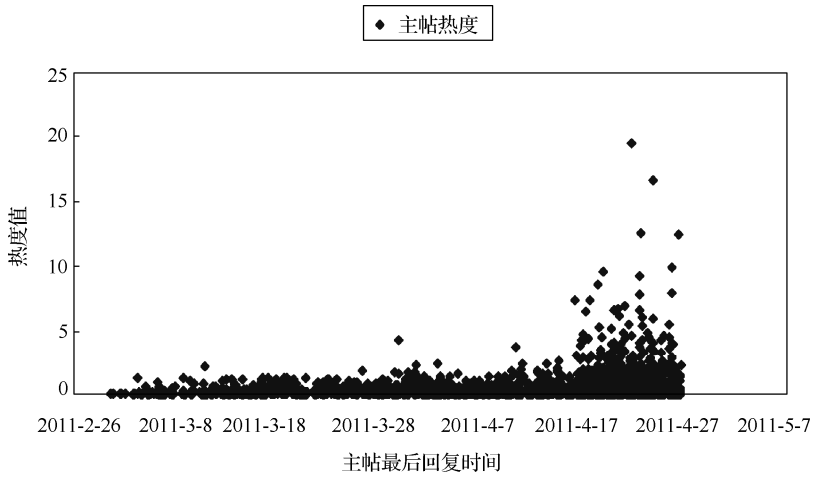


图 5-4　主帖热度的时间分布图

在热帖的突发性特征判断上，以这一时间区间发生的“药家鑫”事件为例，从上述 406 个热帖中提取出与该主题相关的具有突发性特征的主帖共计 6 个。由于“药家鑫”事件本身就是该时间区间的热点事件，以此作为训练样本具有较高的可信度。在实验中，令 $\Delta t = 1$，并随机向样本集中添加 7 个与此主题无关的突发性热帖，根据不同的阈值进行 7 组实验，其实验结果如图 5-6 所示。

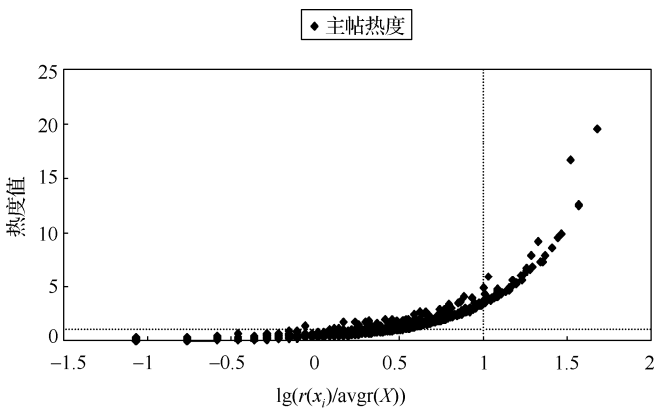


图 5-5　主帖热度在 $\lg(r(x_i)/\mathrm{avgr}(X))$上的分布

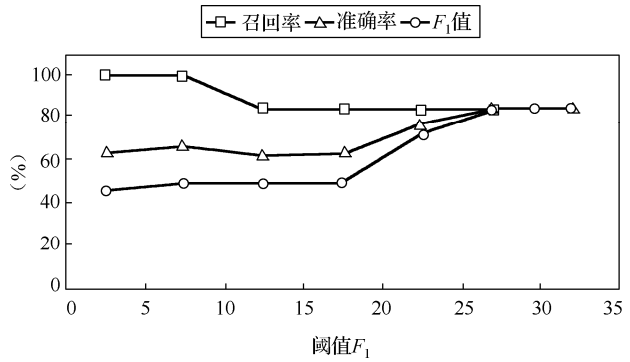


图 5-6　不同阈值下的突发性热帖发现效果

从图 5-6 中可以看出，当 $\theta_1 = 30$ 时，召回率、准确率以及 F_1 值表现都不错。因此在计算回复加速度时取阈值 $\theta_1 = 30$。依据式（5-26）和式（5-27）统计上述 406 个热帖在各时间节点上的回复加速度，并删除不在预定阈值范围内的热帖。通过主帖回复加速度阈值的选择，确定具有突发性特征的主帖有 11 个。这表明通过对候选话题集进行噪声过滤后，原本需要处理的 4248 个帖子，现在只需要处理 11 个帖子。这种过滤方法不仅减小了检测算法的复杂度，同时也提高了检测算法的准确率。对应的 11 个主帖的回复加速度变化情况如图 5-7 所示。

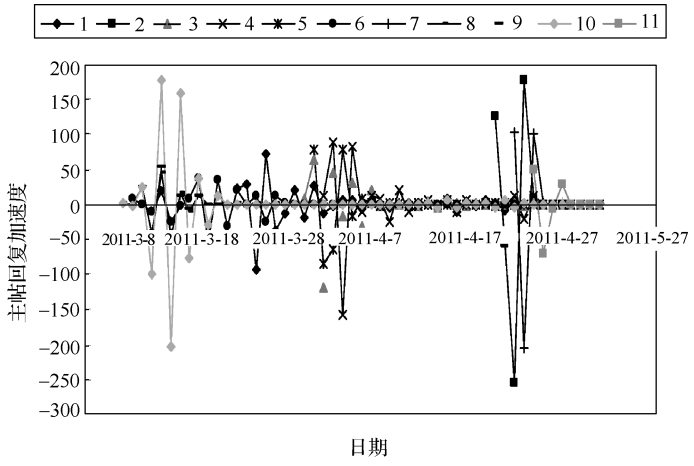


图 5-7　突发性热帖加速度时间变化图

对上述的 11 个主帖标题进行分词，然后采用算法 5-4 进行热点话题检测，得到如表 5-11 所示的检测结果。

表 5-11　突发性热点话题检测结果

突发性热点话题	起始日期	终止日期
中石油给日本地震捐款	2011-3-24	2011-4-21
药家鑫被判死刑	2011-3-31	2011-4-25
新浪微博封杀一剑传媒	2011-4-1	2011-4-19
副乡长抢夺民企	2011-3-12	2011-4-18
蹲监“被死亡”，千万资产被乡领导贱卖据为己有	2011-3-13	2011-4-19

另外，在突发性热点话题跟踪方面，以“药家鑫”突发性热点话题为例，统计6个对应主帖在各个时间点上的回复加速度，计算其平均值，绘制如图 5-8 所示的平均回复加速度变化曲线。通过与实际事件的发展进程进行比对，表明其跟踪效果与实际事件发展基本吻合。

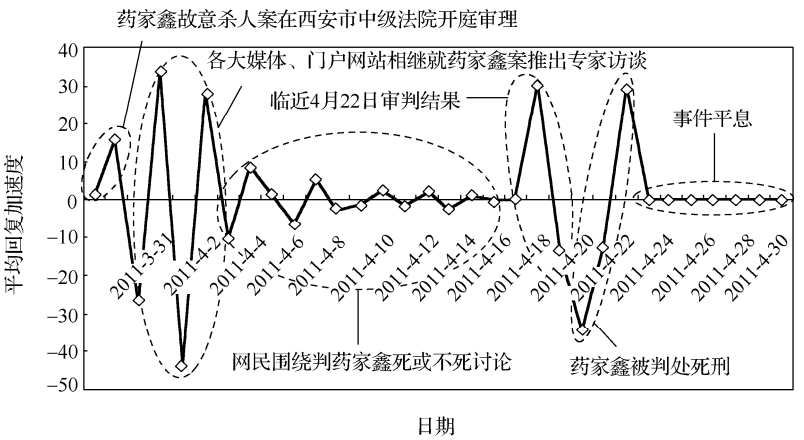


图 5-8　“药家鑫”事件突发性热点话题检测与跟踪结果

第6章

文本分割技术

6.1 引言

文本分割是一种重要的文本分析技术，通过文本分割，自动识别出一个文本中具有独立意义的单元或段落之间的边界，为进一步细化文本分析提供基础。

文本分析主要是对文本的主题或者类别进行分析，通常以文本的自然段落为单位，这样的自然段落往往不能完整地表达子主题，很难对文本的结构做出分析和判断。而文本结构是对文本进行细化处理的基础，很多文本处理任务都是建立在文本结构分析结果的基础上。文本分割为文本结构分析提供了有效手段。

网络舆情分析的对象是海量的非结构化文本信息，在分析之前，需要根据应用需求对非结构化文本信息进行某些预处理。文本分割是一种非结构化文本预处理的重要手段，首先将一个多主题的网页文本按主题分割成若干个文本块，以文本块为单位进行处理，以提高自然语言处理的效果。文本分割技术在文本自动分类、自动摘要、自动问答以及信息检索等自然语言处理应用中发挥重要的作用。例如，在文档自动摘要系统中，按主题进行文本分割后，可以从每个分割单元中抽取出相应的主题信息，再将各个主题信息进行整合，这样更容易实现对原文本的自动摘要任务。在信息检索系统中，以篇章为单位的信息检索结果往往是一篇篇与检索关键字相关的文档，而用户很少有耐心读完整个文档来确定检索结果是否为感兴趣的内容。如果结合文本分割技术，将不再以文档为检索的最小单位，而是以语义段落作为最小的单位，这样用户在检索的时候就能更准确地得到其感兴趣内容的位置，而不需要通读全文，缩小了范围，节省了时间，提高了效率。

文本分割的任务是根据主题相关性找到主题与主题之间的分割边界，将原有的文本内容划分成若干个不同的部分，即分割单元，使得分割单元内部具有最大的主题相关性，而分割单元之间具有最小的主题相关性。因此，对于文本分割方法来说，需要解决的根本问题就是主题相关性度量和边界划分策略。

本章主要介绍基于 LDA 模型和 VSM 模型的文本分割技术。

6.2 基本概念

6.2.1 文本分割点

文本分割的任务是根据主题相关性找到主题与主题之间的分割边界，将原有的文本内容划分成若干个不同的部分，即分割单元，使得分割单元内部具有最大的主题相关性，而分割单元之间具有最小的主题相关性。

在不同的应用中，文本分割点的位置选择可能有所不同，可以选择三种分割点位置：词之间、句子之间以及段落之间。例如，在语音数据流中，由于缺乏句子和段落标记，可以选择词之间作为分割点；在文本数据流中，由于缺乏段落标记，可以选择句子之间作为分割点；如果对一篇自然文本进行分割，可以选择句子之间或者段落之间作为分割点。实际上，不同的分割点位置，对于文本分割技术来说并没有本质上的差别，不同之处主要在于候选分割点位置的预测和选择。

对于三种分割点位置，词和句子的划分相对比较简单，而段落划分要复杂一些。对于文本中的段落划分，主要有以下几种方法：

（1）按照逻辑结构划分。以文本的逻辑结构作为段落划分的依据，文本的逻辑结构包括章节、自然段、句子等，其中保留了文本的逻辑结构信息和语义特征，段落是相对完整的信息单元，保证了信息的相对完整性。然而，这种划分方法存在的主要问题是文本逻辑结构识别往往比较困难，而且逻辑上的段落存在长度上的差异，必须做长度归一化处理。

（2）按照等长自然段划分。把相邻的自然段聚集在一起来确定段落，使每个分块长度大于或等于预设的字节数，同时保证自然段的完整，即分段的边界与自然段边界相一致。这种划分方法主要为了避开按照逻辑结构划分而带来的文本逻辑结构识别和长度差异问题。由于相邻自然段的聚集以长度为依据，而不是文本的逻辑结构和语义信息，因此在一定程度上存在段落和逻辑单元的不匹配问题。

（3）按照话题迁移划分。段落通常是关于某个话题的描述，不同的段落往往描述的是不同的话题，因此可以根据话题的迁移来划分段落。在技术上，以句子或自然段为单位，通过计算和比较相邻句子或自然段的相似度来判断是否出现话题迁移，从而实现对段落的划分。这种划分方法能够保证段落的逻辑和语义，但处理起来比较复杂。

（4）按照固定长度词序列划分。按照固定长度词序列划分段落，又称为“窗口段落”，直接按照相等的词汇量来划分段落，第一个段落的起始位置可以由文本中出现的第一个检索词的位置来确定，后续每个段落划分的起点是上个段落的末点。这种划分方法的操作简单，也避免了长度归一化问题，但它完全忽略了文本的逻辑结构和语义特征，划分质量受到一定的影响。

（5）按照文本语义划分。按照文本中的语义，提取语义相近的自然段落作为一个段落划分，这种段落称为语义段落或语义段，每个语义段描述了一个相对独立的主题，这里的主题是指介于文本与自然段落之间的一个语言单位，一个主题表达一个相对独立的话题，形式上由文本中的若干个自然段组成，各个主题相连构成整个文本，因此文本中的主题划分可以认为是文本段落的提取过程。这种划分方法比较适合于对文本进行细粒度分析。

不同的段落划分方法各有优缺点，在实际应用中，需要根据具体的应用需求来确定选择哪一种方法。例如，在文本情感分析中，可以选择文本语义划分方法，以便对文本情感做细粒度的分析。

6.2.2　文本分割方法

新闻报道切分是 TDT 任务之一，其目标是将一个语言信息流分割为不同的独立新闻报道。因此 TDT 研究推动了文本分割技术的发展，提出了多种文本分割方法，这些方法总体上可以分为 4 类：基于词聚集的方法、基于语言特征的方法、基于统计的方法以及其他方法。

1．基于词聚集的方法

基于词聚集的方法假定相似或相关的词倾向于出现在同一主题段落内，文本内部有机的组织是文本的一个重要特征，一个任意的句子集合并没有这种特征，而篇章内部的紧凑性是使得文本篇章成为有机组织的一个重要因素。篇章内部的紧凑性表现为：篇章的一个元素（词、术语或句子）能够在同一篇文章的上下文中得到诠释。

典型的基于词聚集的方法有 Text Tiling 算法、Lexical Chains 方法、Dotplotting 方法以及 LCP（Lexical Coherence Profile）方法等。这些方法都要求计算词之间的相似性或相关性，但如何计算和量化词之间的相关性是这些方法的难题。例如，Dotplotting 方法通过直接统计一个上下文中相同的单词数量来得到词密度信息，并用最大化词密度信息作为文本主题边界分割的依据。

2．基于语言特征的方法

基于语言特征的分割方法是指利用某种策略从语料库中提取词特征或者韵律特征，通过分析它们与主题段落首尾的关系来确定段落边界。这种方法一般用于特定文本类或者语音流的处理。与书面文本分割相比，语音流分割存在更大的难度，其原因是书面文本有一些形式化的信息，如标题、段落等，可以辅助文本的分割。而语音流不仅没有这些信息，并且语音识别的结果还会存在一定的错误率，书面文本分割方法在分割语音流时效果不佳。算法性能（如准确率、召回率等）在很大程度上取决于语料库中语料的选择、特定的说话人、所使用的学习策略等。

基于语言特征的分割方法主要有：

（1）决策树方法。利用决策树对语料库进行分析，挖掘声学-韵律与主题变换的关系。这种方法的优点是可以实现自动编码，不需要手动编码。

（2）隐马尔可夫模型（HMM）。利用 HMM 对主题段落的开始、中间及结束句子建模分析，所使用的特征除了“标志词”外，还有句子长度、连续聚类倾向以及词两次出现间的距离等。这种方法的有效性主要依赖于对主题段落的首尾边界有暗示作用的词，虽然也考虑其他语言特征，但这些特征所发挥的作用比较小。

（3）综合方法。综合利用 HMM 和决策树两种方法，利用词和韵律信息对语音流进行分割，词信息通过语音识别获得，韵律信息从语音波形上直接提取，主要有持续时间特征及音质特征。实际上，韵律模型本身就可以达到基于词的分割效果，如果将两者结合起来，则有助于降低识别错误率。

3．基于统计的方法

基于统计的方法是建立一个统计语言模型进行文本分割，利用统计语言模型（如指数模型）在标注好主题边界的训练语料中抽取一些能够指示边界的特征，该模型使用了两类特征：主题性特征和提示词特征。

4．其他方法

其他文本分割方法还有 LSA（Latent Semantic Analysis）方法、动态规划（Dynamic Programming）方法、局部内容分析（Local Context Analysis）方法以及侧面隐马尔可夫模型（Aspect Hidden Markov Model）等。

文本分割中存在的主要问题如下：

（1）段落长短问题。在一篇文章中，作者通常会根据自己行文的需要选择段落的长短。但在文本分类中，通常将各个待标记的文本看成对等的个体，这就意味着对每一个文本赋予同样的权重，在标记结果生成时，如果存在标记错误的文本，则根据它们的数量来计算其错误率，而不需要考虑文本的长度。但在文本分割中情况就不同了，当一篇较长的段落被错分时，显然不能与一个较短的段落被错分同等对待，较长的段落在分割中应该占有更高的权重，而且当这种段落被错分时应该给予更高的惩罚，因此如何平衡长短段落之间的关系也是文本分割的重要问题。比较流行的做法是将一个自然段落进行细分，在自然段落中设置一个长度，将自然段落中的词或特征按照这个长度切分成若干个标记序列，每一个标记序列是对等的对象。当一个自然段落较长时，标记序列的数量就会比较多，反之，数量就会比较少。这样的标记序列具有表征类别大小和相互对等的特性，可以和分类中的一篇文本相对应，对于标记序列较多的段落，在文本分割中自然占有更重要的位置。

（2）体裁问题。文本一般含有多种体裁，对于不同的体裁，其写作方式也不固定、不统一，这就给文本分割带来了一定的难度。在实际处理时，通常将文本分割的对象限制在一定的范围内，以常用的文本体裁为对象体裁，如记叙文、说明文、议论文等，这样的文章一

般有主题明确、段落清晰、结构严谨等特点，具有一定的代表性和实用性。

（3）子主题跳转问题。一篇文章往往由一个核心主题和若干子主题组成，核心主题确定了一篇文章的讨论范围和文章架构，在核心主题的基础上，作者会根据其所要讨论题目的具体情况选用例证、引用、对比等手法，形成相应的子主题。在文本分割中，这种子主题的跳转一般不具有客观的评价标准，并且在两个相邻的子主题之间往往既存在着相似性，也存在着区别性。应当根据区别程度的大小，将这两个相邻的段落划分开，但其中的阈值很难确定。除此之外，相似度计算方法也是一个确定分割位置的决定性因素，不同的相似度计算方法必然会产生不同的分割结果，究竟哪种相似度计算方法更适合于文本分割？采用统一的相似度计算方法能否适用于所有文本？这些问题都是文本分割技术所要解决的问题。

6.2.3　文本分割算法评价

对于文本分割算法性能的评价，主要存在两个问题。第一，评价标准缺乏客观性，人们对段落边界的位置以及文本分析的粒度往往难以形成一致的看法和观点，这就给文本分割结果的评价增加了难度。通常采用大多数人的意见作为评价标准。第二，不同的应用对文本分割的准确性有不同的要求，对于文本分割准确性的评价标准也应根据应用要求而异。比较合理的评价方法应当包括直接评价和间接评价两种方式，直接评价是指采用某种评价标准对算法的边界判断能力进行直接评价，而间接评价则是指将算法嵌入各自不同的应用中进行评价。由于间接评价手段差异较大，因此下面主要介绍几种常用的直接评价指标。

1．准确率和召回率

准确率和召回率是文本处理中常用的评价指标，在文本分割中，准确率和召回率定义如下。

（1）准确率（P）：系统正确识别的边界数量与系统返回的边界数量之比，计算公式为 $P = A/(A+B)$，其中，A 为系统正确识别的边界数量，B 为系统错误识别的边界数量。

（2）召回率（R）：系统正确识别的边界数量与语料库中所有的边界数量之比，计算公式为 $R = A/(A+C)$，其中，A 为系统正确识别的边界数量，C 为系统未识别的边界数量。

使用准确率和召回率指标来评价文本分割的效果时存在两个缺陷：一是准确率和召回率之间存在一个权衡问题，提高准确率是以降低召回率为代价，反之亦然。二是准确率和召回率主要考虑的是绝对匹配的结果，实际上，距离正确分割点较近的错误分割点比距离较远的错误分割点的性能要好，而准确率和召回率则一视同仁，无法体现出这种差别。

2．F_1 度量

为了在准确率和召回率之间找到平衡点，更好地评估算法性能，可以采用一种综合评价指标 F，F 值计算公式如下：

$$F=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R} \tag{6-1}$$

式中，P 为准确率，R 为召回率，β 为调整查准确率和召回率在评价函数中所占比重的参数，通常取 $\beta=1$，这时的评价指标变为：

$$F_1=\frac{2PR}{P+R} \tag{6-2}$$

$\beta=1$ 的 F 值记为 F_1，其值越高，其综合性能越好。

3．P_k 度量

P_μ 度量是一种基于错误罚分的评价指标，以克服准确率和召回率评价指标存在的不足。设 ref、hyp 为包含 n 个句子的语料的两个分割，则 P_μ 定义如下:

$$P_\mu(\text{ref},\text{hyp})=\sum_{1\leqslant i\leqslant j\leqslant n} D_\mu(i,j)\delta_{\text{ref}}(i,j)\overline{\oplus}\delta_{\text{hyp}}(i,j) \tag{6-3}$$

式中，δ_{ref} 是指示函数，若其值为 1，则由参数指定的两个索引属于同一文档；若其值为 0 则分属不同文档。类似地，如果两个索引属于同一文档，则 $\delta_{\text{hyp}}=1$，否则 $\delta_{\text{hyp}}=0$。$\overline{\oplus}$ 是运算符，对左右两个函数的真假值进行异或运算。函数 D 为距离概率分布，用于体现从语料库中随机抽取的两个句子间的距离，其值很大程度上依赖于文档平均间距的某些参数。

P_μ 的导出相当复杂，可以将其简化为 P_k，即令 $D=D_k$，使得距离取固定值 k，其含义为随机抽取间隔 k 个词的词对，判断它们属于同一段落或者分属不同段落的概率。P_k 首先将 k 设置为真实段落长度平均值的一半，然后通过移动长度为 k 的窗口计算罚分。P_k 的计算公式如下：

$$P_k=P(\text{seg})P(\text{miss})+(1-P(\text{seg}))P(\text{false alarm}) \tag{6-4}$$

式中，P(seg)是指距离为 k 的两个句子分属不同主题段落的概率，而 $1-P$(seg)是指距离为 k 的两个句子属于同一主题段落的概率，将两个先验概率取等值，即 P(seg) = 0.5。P(miss)是算法分割结果缺少一个段落的概率，P(false, alarm)是算法分割结果添加一个段落的概率。

P_k 的取值范围为 0～1，如果一个算法判断的边界完全正确，则 P_k 取值为 0。该方法体现了判断边界与真实边界之间的距离对于算法优劣的影响，距离越大则罚分越高。

研究表明，P_k 度量通过罚分来评价算法识别率存在一定的局限性，在某些情况下存在不公平现象，影响它的客观性。

4．WindowDiff 度量

WindowDiff 度量对 P_k 度量进行了改进，WindowDiff 定义如下：

$$\text{WindowDiff}(\text{ref},\text{hyp})=\frac{1}{N-K}\sum_{i=1}^{N-K}(\left|b(\text{ref}_i,\text{ref}_{i+k})-b(\text{hyp}_i,\text{hyp}_{i+k})\right|>0) \tag{6-5}$$

式中，$b(i, j)$表示整句 s_i 和整句 s_j 间的边界数量，N 表示文本中的整句数量，k 取真实段落平均长度的一半，ref 代表真实分割，hyp 代表算法分割。

WindowDiff 也是在 0 到 1 之间取值，WindowDiff 值越小，算法分割的效果越好。

6.3　基于 LDA 模型的文本分割

前面提到，文本主题相关性分析是文本分割中必须解决的主要问题。因此，需要采用某种适当的表示模型来描述文本主题，以便对主题的相关性进行度量和计算。常用的表示模型有向量空间模型和语言模型，其中，LDA（Latent Dirichlet Allocation）模型是一种效果比较好的语言模型。

下面介绍 LDA 模型及其在文本分割中的应用。

6.3.1　LDA 模型

1．LDA 模型概念

LDA 模型是一种语义模型，与传统的语义模型相比，LDA 模型是一个多层的产生式概率模型，包含词、主题和文本（文档）等三层结构，更符合实际的文本特点。LDA 模型将每个文本表示为一个主题混合体，它假设词是由一个主题混合产生的，每个主题是固定词汇表中的一个多项式分布，这些主题被文本集合中的所有文本所共享，每个文本有一个特定的主题比例，从狄利克雷（Dirichlet）分布中抽样产生。

在 LDA 模型中，定义如下术语：

（1）词 w 是离散数据的基本单位，由 $\{1, 2, \cdots, v\}$ 进行索引，例如 w_i 表示所给文本中的第 i 个词。

（2）文本 d 是 N 个词的序列，用 $d = (w_1, w_2, \cdots, w_n)$表示，$w_n$ 是序列中的第 n 个词。

（3）文本集 D 由 M 个文本 d 组成，用 $D = (d_1, d_2, \cdots, d_M)$表示。

（4）Dirichlet 分布是一种指数分布，具有有限的维数，便于有效的统计，并且它共轭于多项式分布。它的这些特点便于 LDA 的参数估计和推导。

LDA 模型采用如下的产生式过程来表示一个文本。

（1）产生文本长度 N，它服从泊松分布 Poisson(ξ)，泊松分布是一个离散分布，适合描述单位时间内随机事件发生的次数。N 服从泊松分布并不是关键，可以替换成其他的离散分布。

（2）产生参数 θ，它是一个 k 维向量，服从 Dirichlet 分布 Dir(α)，Dirichlet 分布是一个连续多随机变量分布。

（3）对于文本中的每一个词 w_n，有如下产生过程：

a）产生一个主题 z_n，它服从二项分布 Multinormial(θ);

b）产生一个词 w_n，它服从二项分布的条件概率 $p(w_n|z_n,\beta)$，其中 z_n 是上一步产生的主题。

其中，$\boldsymbol{\beta}$ 是一个 $k\times V$ 的矩阵，k 是主题数，V 是词数。每个词 w 都表示成一个 V 维向量，其中只有一个元素值为 1，其他都为 0。$\boldsymbol{\beta}$ 矩阵中的值 β_{ij} 表示词 j 在主题 i 中出现的概率。给定一个主题 z_n 和 $\boldsymbol{\beta}$ 矩阵，实际上就是取矩阵的一行，该行就是某主题下的词分布，根据这个分布产生一个词。

LDA 模型如图 6-1 所示，白色圆表示潜在变量，深色圆表示可观察值，矩形表示重复过程，大矩形表示从 Dirichlet 分布中为文本集中的每个文本 d 反复抽取主题分布θ，小矩形表示从主题分布中反复抽样产生文本 d 的词{w_1，w_2，…，w_N}。

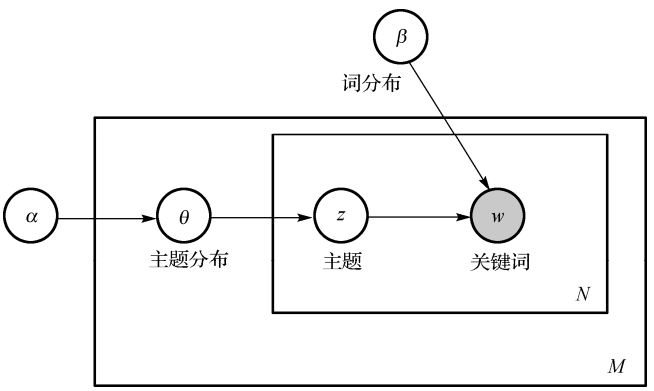


图 6-1　LDA 模型

从图 6-1 可以看出，LDA 模型是一个三层生成的概率模型，z 和 w 是词层次的变量，对于每个文本的每个词都要采样一次；θ 是文本层次的变量，对每个文本都要采样一次；α 和 β 是文本集层次的参数，在文本集生成过程中只采样一次。α 参数可理解为，在从文本观察到任何实际数据之前，α 为一个文本中主题 j 抽样次数的先验观察计数。对于主题分布θ，设置一个 Dirichlet 先验分布能够平滑主题分布，平滑的程度由参数 α 决定。

在 LDA 模型中，需要解决一个关键问题是推理问题。对于给定文本，采用下式计算潜在变量的后验分布概率：

$$P(\theta,z|\ w,\alpha,\beta)=\frac{p(\theta,z,w|\ \alpha,\beta)}{p(w|\ \alpha,\beta)} \tag{6-6}$$

这个分布概率很难计算，通常采用近似推理方法来估计主题 z 的后验概率分布，Gibbs 抽样算法是常用的近似推理方法。

假设有 T 个主题，则文本中的第 i 个词 w_i 可以表示如下：

$$P(w_i)=\sum_{j=1}^{T}p(w_i|\ z_i=j)p(z_i=j) \tag{6-7}$$

式中，z_i 为潜在变量，表明第 i 个词 w_i 取自该主题，$P(w_i|z_i=j)$为词 w_i 属于主题 j 的概率，$P(z_i=j)$为主题 j 属于当前文本的概率，$P(z)$为在特定文本中主题 z 的概率分布，$P(w|z)$表示在给定主题 z 的情况下，词 w 的概率分布。

对于文本中的每个词 w_i，采用如下生成过程：首先根据主题分布抽样一个主题，然后根据相应的主题-词分布来选择 w_i。

为了简化符号，通过设置一个对称 Dirichlet(β)先验分布 φ 表示词分布，令 $\varphi = P(w|z=j)$，表示对于主题 j，w 个词上的多项分布；$\theta=P(z=j)$表示对于文本 d，T 个主题上的多项分布。

这样，扩展后的 LDA 模型如图 6-2 所示，其中，β 可理解为在观察到文本集中的任何词之前，在一个主题上词抽样次数的先验观察计数。对于词分布 φ，设置一个 Dirichlet 先验分布来平滑词分布，平滑的程度由参数 β 决定。

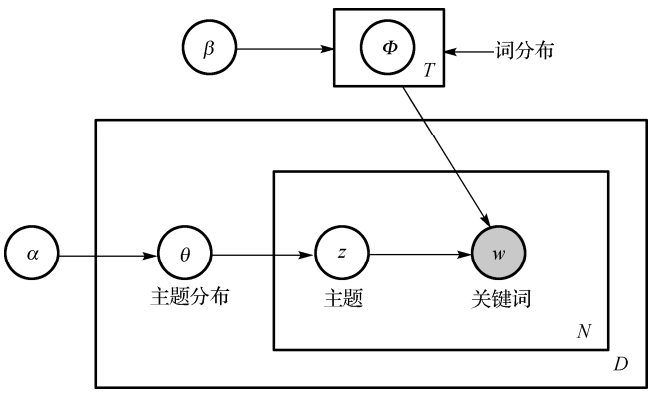


图 6-2　LDA 模型的扩展

在图 6-2 中，白色圆和深色圆分别表示观察到的变量和潜在变量，Φ、θ 和 z 是需要推导的变量。图中箭头表示变量直接的条件依赖，矩形表示重复抽样，包含变量 z 和 w 的小矩形表示主题和词的重复抽样，直到生成了文本 d 的 N_d 个词。包含变量 θ 的大矩形表示对于每个文本 d 抽样主题分布概率，直到 D 个文本处理完。变量Φ表示主题层次，包含变量Φ的小矩形表示对于每个主题 z 要重复抽样词分布的概率，直到 T 个主题全部生成。α 和 β 作为常量，实验表明，选取 $\alpha=50/T$，$\beta=0.01$ 是比较合适的。

2．Gibbs 抽样算法

1）Gibbs 抽样算法描述

为了简化词的概率分布，利用 Gibbs 抽样间接求得 θ 和Φ的值，而不是直接计算 θ 和Φ的值。MCMC（Markov Chain Monte Carlo）是一套从复杂的概率分布中抽取样本值的近似

迭代方法，Gibbs 抽样作为 MCMC 的一种简单实现形式，其目的是构造收敛于某个目标概率分布的马尔可夫链，并从马尔可夫链中抽取接近于该概率分布值的样本。因此，目标概率分布函数的选取便成为 Gibbs 抽样的关键。

Gibbs 抽样依次抽取文本集中的每个词，估计 $P(z_i)$，z_i 为潜在变量，表明第 i 个词 w 取自该主题。条件依赖于其他 z_{-i}；依据这个条件分布，相对应的主题被抽样出来，并被保存为 $z_i=j$。后验概率为 $P(z_i=j|z_{-i}, w_i, d_i, \cdot)$， 其中，$z_i=j$ 表示将词 w 分配给主题 j，z_{-i} 表示所有 $z_k(k\neq i)$的分配，“$\cdot$”符号表示所有其他已知或观察到的信息，如其他词和文本 w_i。其计算公式如下：

$$P(z_i=j|\ z_{-i},w_i,d_i,\cdot)=\frac{\dfrac{C_{w,j}^{\mathrm{WT}}+\beta}{\sum_{w=1}^{W}C_{w,j}^{\mathrm{WT}}+W\beta}\times\dfrac{C_{d,j}^{\mathrm{DT}}+\alpha}{\sum_{t=1}^{T}C_{d,t}^{\mathrm{DT}}+T\alpha}}{\dfrac{C_{w,j}^{\mathrm{WT}}+\beta}{\sum_{t=1}^{T}\sum_{w=1}^{W}C_{w,j}^{\mathrm{WT}}+W\beta}\times\dfrac{C_{d,j}^{\mathrm{DT}}+\alpha}{\sum_{t=1}^{T}C_{d,t}^{\mathrm{DT}}+T\alpha}} \tag{6-8}$$

式中，$\boldsymbol{C}^{\mathrm{WT}}$ 为词-主题计数矩阵，维数 $=W\times T$；$\boldsymbol{C}^{\mathrm{DT}}$ 分为主题-文本计数矩阵，维数$=D\times T$；$C_{w,j}^{\mathrm{WT}}$ 为分配给主题 j 与 w 相同的词数（即词的数量，下同）；$\sum_{w=1}^{W}C_{w,j}^{\mathrm{WT}}$ 为分配给主题 j 的所有词数；$C_{d,j}^{\mathrm{DT}}$ 为文本 d 中分配给主题 j 的词数；$\sum_{t=1}^{T}C_{d,t}^{\mathrm{DT}}$ 为 d 中所有被分配了主题的词数，所有的词数均不包括此次 $z_i=j$ 的分配。

3．Gibbs 抽样过程

Gibbs 抽样过程如下：

（1）z_i 初始化为 1 到 T 之间的某个随机整数，作为马尔可夫链的初始状态，i 从 1 循环到 N，N 是语料库中所有出现在文本中的词数。

（2）i 从 1 循环到 N，根据式（6-8）将词分配给主题，获取马尔可夫链的下一个状态。

（3）第（2）步迭代足够次数后，认为马尔可夫链接近于目标分布，取 z_i（i 从 1 循环到 N）的当前值作为样本记录下来。为了保证自相关较小，每迭代到一定的次数，记录其他样本。

这样，Gibbs 抽样算法可以直接估计 z_i 值。

在一些应用中还需要估计 Φ 和 θ 值，它们的估算公式如下：

$$\Phi_i'^{(j)}=\frac{C_{ij}^{\mathrm{WT}}+\beta}{\sum_{k=1}^{W}C_{kj}^{\mathrm{WT}}+W\beta}\qquad \theta_j'^{(d)}=\frac{C_{dj}^{\mathrm{DT}}+\alpha}{\sum_{k=1}^{T}C_{dk}^{\mathrm{DT}}+T\alpha} \tag{6-9}$$

式中，C_{ij}^{WT} 表示词 i 被分配给主题 j 的次数，$\sum_{k=1}^{W}C_{kj}^{\mathrm{WT}}$ 表示分配给主题 j 的所有词数，C_{dj}^{DT} 表示文本 d 中分配给主题 j 的词数，$\sum_{k=1}^{T}C_{dk}^{\mathrm{DT}}$ 表示文本 d 所有分配给主题的词数。

6.3.2　LDA 模型改进

LDA 模型主要完成对文本的模型化，在此基础上还需要结合相似度度量方法和边界识别策略才能完成文本分割任务。如果 LDA 模型在模型化文本的同时，还能完成文本分割任务，无疑会提高文本分割的效率。这就需要对 LDA 模型进行改进，改进后的 LDA 模型称为 LDA-I（Latent Dirichlet Allocation-Improved）模型。下面介绍 LDA-I 模型。

假设有一个文本集，文本集中有 U 个文本，第 u 个文本包含 N_u 个词，词来源于大小为 W 的词汇表。属于第 u 个文本的词由 w_u 表示，由 $w_{u,i}$ 进行索引。

LDA 模型将每个文本表示为一个主题混合体，每个主题是固定词汇表上的一个多项式分布。假定主题分布符合马尔科夫结构：在大概率条件下，第 u 个文本与第 u–1 个文本具有相同的主题分布概率，否则就抽样一个新的主题分布概率。为了表示这种相关性，为每个文本引入了一个二进制开关变量 c，表示第 u 个文本是否与第 u–1 个文本具有相同的主题分布。为了叙述方便，将 u 称为一个文本，实质上 u 是一个抽象的概念，可以表示一个文本、一个文本的段落，甚至是文本的单个句子。

这里所关注的问题是，在文本集中哪个文本或哪些文本的主题分布概率发生了改变。为了形式化，引入下面三个符号：

（1）用符号 c_u 表示在第 u 个文本中主题分布概率是否发生了改变。

（2）用符号 π 表示 $c_u=1$ 的先验分布概率，令 $P(c_u=1)=\pi$，且 $\pi\sim\mathrm{Beta}(\gamma)$。

（3）用符号 $\theta^{(u)}$ 表示第 u 个文本的主题概率分布，$\theta_t^{(u)}$ 表示主题 t 的概率。设定：如果 $c_u=0$，则 $\theta^{(u)}=\theta^{(u-1)}$；否则 $\theta^{(u)}\sim\mathrm{Dir}(\alpha)$。

通过引入三个符号，对 LDA 模型做进一步的扩展和改进，改进后的 LDA 模型称为 LDA-I 模型，如图 6-3 所示。与图 6-2 相比，在图 6-3 中增加了 3 个符号：γ、π 和 c_u。γ 和 π 作用类似于 α 和 β，也被视为常量；符号 c_u 表示在第 u 个文本中主题分布概率是否发生了改变，包含 c_u 的矩形表示对于每个文本 u 抽样的分布概率，直到 U 个文本处理完。

根据贝叶斯准则，得到如下公式：

$$P(z,c|\ w)=\frac{P(w|\ z)P(z|\ c)P(c)}{\sum_{z,c}P(w|\ z)P(z|\ c)P(c)} \tag{6-10}$$

式中，$P(w|z)$为变量 Φ 的积分，$P(z|c)$为变量 θ 的积分，$P(c)$为变量 π 的积分。分子可以通过积分运算得到，而分母仍需要使用 Gibbs 抽样求得后验概率分布 $P(z, c|w)$。

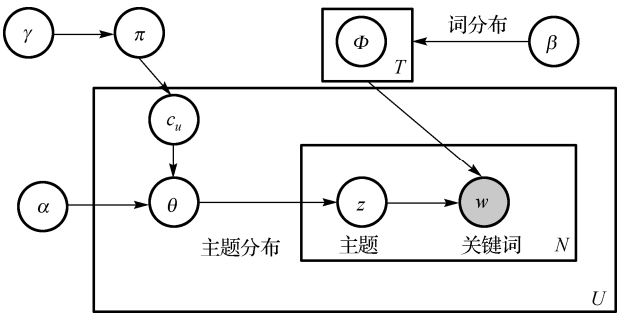


图 6-3　LDA-I 模型

对于每个词，使用 Gibbs 抽样出主题分配 $z_{u,i}$，其条件依赖于所有其他主题分配 $z_{-(u,i)}$、所有主题变化标识 c 和所有的词 w。然后为每个文本抽样出主题变化标识 c_u，其条件依赖于所有其他主题变化标识 c_{-u}、所有的主题分配 z 和所有的词 w。因此得到如下公式：

$$P(z_{u,i}\mid z_{-(u,i)},c,w)=\frac{n^{(t)}_{w_{u,i}}+\beta n^{(s_u)}_{z_{u,i}}+\alpha}{n.^{(t)}+W\beta n.^{(s_u)}+T\alpha} \tag{6-11}$$

式中，变量 c_u 的条件概率表示文本 u 是一个新段落的概率。

在抽样 c_u 的过程中，合并或者分割段落。因此得到如下表达式：

$$P(c_u\mid c_{-u},z,w)\propto\begin{cases}\dfrac{\prod_{t=1}^{T}\Gamma(n_t^{(S_u^0)}+\alpha)}{\Gamma(n_\cdot^{(S_u^0)}+T\alpha)}\dfrac{n_0+\gamma}{N+2\gamma} & c_u=0\\ \dfrac{\Gamma(T\alpha)}{\Gamma(\alpha)}\dfrac{\prod_{t=1}^{T}\Gamma(n_t^{(S_{u-1}^1)}+\alpha)}{\Gamma(n_\cdot^{(S_u^1)}+T\alpha)}\dfrac{\prod_{t=1}^{T}\Gamma(n_t^{(S_{u-1}^1)}+\alpha)}{\Gamma(n_\cdot^{(S_u^1)}+T\alpha)}\dfrac{n_1+\gamma}{N+2\gamma} & c_u=1\end{cases} \tag{6-12}$$

式中，S_u^1 表示当 $c_u=1$ 的 S_u，S_u^0 表示当 $c_u=0$ 的 S_u，对于参数 α、β 和 γ，令 $\alpha=50/T$，$\beta=\gamma=0.01$。

LDA-I 模型的算法流程如图 6-4 所示。

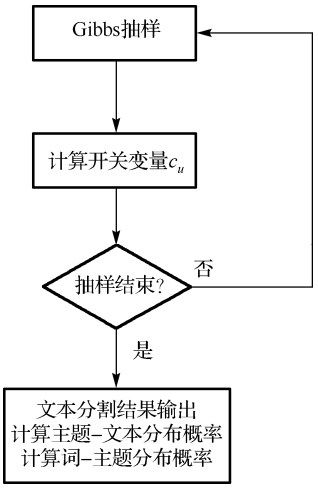


图 6-4　LDA-I 模型算法流程

6.3.3　相似度计算

在基于主题分布概率 θ 计算句子间的相似度时，需要选择合适的相似度度量方法，常用的相似度度量方法有如下几种。

（1）余弦系数。计算公式如下：

$$\mathrm{Sim}_{\mathrm{Cos}}=\frac{\sum_{i=1}^{T}\theta_i^{(d_1)}\times\theta_i^{(d_2)}}{\sqrt{\sum_{i=1}^{T}(\theta_i^{(d_1)})^2\times\sum_{i=1}^{T}(\theta_i^{(d_2)})^2}} \tag{6-13}$$

式中，$\theta^{(d)}$ 为主题分布概率，d 为主题数目，T 为主题数量，下同。

（2）Jaccard 系数。计算公式如下：

$$\mathrm{Sim}_{\mathrm{Ja}}=\frac{\sum_{i=1}^{T}(\theta_i^{(d_1)}\times\theta_i^{(d_2)})}{\sum_{i=1}^{T}(\theta_i^{(d_1)})^2+\sum_{i=1}^{T}(\theta_i^{(d_2)})^2+\sum_{i=1}^{T}\theta_i^{(d_1)}\times\theta_i^{(d_2)}} \tag{6-14}$$

（3）L1 距离。计算公式如下：

$$\mathrm{Sim}_{\mathrm{L1}}=1-\frac{\sum_{i=1}^{T}|\ \theta_i^{(d_1)}-\theta_i^{(d_2)}|}{2} \tag{6-15}$$

（4）Hellinger 距离。计算公式如下：

$$\mathrm{Sim}_{\mathrm{Hel}}=\sum_{i=1}^{T}\sqrt{\theta_i^{(d_1)}\times\theta_i^{(d_2)}} \tag{6-16}$$

（5）Clarity 系数。计算公式如下：

$$\mathrm{Sim}_{\mathrm{KL}}=\frac{1}{2}\left[D(\theta^{(d_1)})+D(\theta^{(d_2)})\right] \tag{6-17}$$

（6）Jensen-Shannon 发散。计算公式如下：

$$\mathrm{Sim}_{\mathrm{JS}}=\frac{1}{2}\left[D\left(\theta^{(d_1)},\frac{(\theta^{(d_1)}+\theta^{(d_2)})}{2}\right)+D\left(\theta^{(d_2)},\frac{(\theta^{(d_1)}+\theta^{(d_2)})}{2}\right)\right] \tag{6-18}$$

式（6-18）基于这样一种思想：如果 $\theta^{(d_1)}=\theta^{(d_2)}$，则 $\theta^{(d_1)}=\theta^{(d_2)}=\frac{(\theta^{(d_1)}+\theta^{(d_2)})}{2}$。

6.3.4 边界识别策略

边界识别策略是指在文本分割时对文本边界所采用的估计方法，不同的边界估计策略将对文本分割质量产生一定的影响。在文本分割中，常用的边界估计策略如下。

（1）常数法。设定常数 ξ，如果句子 s_1, s_2 之间的相似度值 $\mathrm{Sim}(s_1, s_2)<\xi$，则认为 s_1, s_2 分属于不同的段落。该方法易于实现。如果 ξ 值选择合适，则分割错误率会比较低。常数法虽然简单，但需要人为设定 ξ 值，最佳值选择往往比较困难。

（2）动态常数法。根据相邻句子间的相似度值表动态改变 ξ 值，假设待分割文本有 n 个整句，则相邻句子间的相似度值表为 $\mathrm{SimTable}=\{\mathrm{Sim}_1, \mathrm{Sim}_2, \cdots, \mathrm{Sim}_{n-1}\}$，其中，$\mathrm{Sim}_i=\mathrm{Sim}\ (s_i, s_{i+1})$，$1\leqslant i\leqslant n-1$，令 $\mathrm{avgSim}=\frac{\mathrm{Sim}_1+\mathrm{Sim}_2+\cdots+\mathrm{Sim}_{n-1}}{n-1}$，$\mathrm{avgmSim}=\frac{(\mathrm{Sim}_2-\mathrm{Sim}_1)+\cdots+(\mathrm{Sim}_{n-1}-\mathrm{Sim}_{n-2})}{n-2}$，如果 $\mathrm{avgmSim}\leqslant\mathrm{Sim}\ (s_1, s_2)\leqslant\mathrm{avgSim}$，则认为 s_1, s_2 分属于不同的段落。

（3）局部最小值法。在相邻句子间的相似值表 SimTable 中选择局部最小值 $\mathrm{Sim}_{\mathrm{mim}}(s_1, s_2)$，首先从每一个局部最小值出发，向左、向右分别寻找距离最近的较大值 $\mathrm{Sim}_{\mathrm{maxl}}$、$\mathrm{Sim}_{\mathrm{max}r}$，然后计算相对深度：$d_{\mathrm{rel}}=\frac{\mathrm{Sim}_{\max}+\mathrm{Sim}_{\max r}}{2\mathrm{Sim}_{\min}(s_1,s_2)}-1$；令 ξ 为一个常数，如果相对深度 $d_{\mathrm{rel}}(s_1, s_2)>\xi$，则 s_1, s_2 分属于不同的段落。

（4）动态规划法。将某个文本分割为 K 个段落 $\{t_1, t_2, \cdots, t_k, \cdots\}$，段落 $t_k\ (1\leqslant k\leqslant K)$ 由首尾句 s_1, s_2 决定，或者 $t_k=[i, j]$，$k\leqslant K$），t_k 的平均相似度值为 $\beta(t_k)=\frac{\sum_{i\in t_k}\sum_{j\in t_k}\mathrm{Sim}(s_i,s_j)}{(j-i+1)^2}$。

6.3.5　算法验证

下面通过实验数据对文本分割算法性能进行测试和验证。

1．实验数据集

1）数据集要求

在文本分割算法测试中，只需要构造一个测试数据集，测试数据集是从互联网中下载的中文语料库，为了能够体现不同的分割算法性能的差异性，在选取语料时有如下的要求：

（1）长度要求。作为测试文本，应该具有一定的长度，使之能够反映出文本的统计特征。按照统计学的观点，当样本数量过少时，无法获得应有的统计信息，其统计结果的偏差较大或置信度不高。类似地，过短的文本也不能获得用于文本分割的统计特性，其实验结果偏差较大或置信度不高。

（2）体裁要求。不同的文章体裁与不同的写作方式和风格相对应，对文本中的特征分布有较大的影响。没有任何一种统计方法是通用的，必须根据具体的问题选用适当的统计方法。对于不同的文章体裁，应当设计相应的分割策略。例如，适用于叙事文和说明文的分割方法，则不适合对散文、诗歌等体裁的文章进行分割。因此有必要对文章体裁加以限制，以确定实验中所要处理的对象。

（3）内容要求。为了测试文本分割算法的性能，一个文本的内容最好围绕一个主题来叙述，并且应包括有若干个子主题，这样的文本才能体现文本分割的意义。

2）数据集构造

根据上述的规则，构造了一个由 400 篇中文语料组成的实验数据集，语料来源于人民日报电子版，内容较为广泛，涵盖了科技说明文、人物传记、时事评论等体裁。为了方便实验，将语料库划分为 4 个测试数据集 $T_{3\text{-}11}$、$T_{3\text{-}7}$、$T_{8\text{-}12}$、$T_{11\text{-}15}$，其中，$T_{x\text{-}y}$ 表示所含主题段落的句子数在 x 和 y 之间。每个测试集都包含有若干个伪文本，作为干扰文本，伪文本是使用不同类的文本连接而成的形式上的文本。每个测试数据集构造完成后，通过人工阅读，指出确切的语义段落分割，并生成答案。测试数据集构成情况如表 6-1 所示。

表 6-1　实验中测试数据集的构成

测　试　集	句　子　数	每个主题段落句子数	测试集平均主题数	伪 文 本 数
$T_{3\text{-}11}$	2 398	3～11	113	100
$T_{3\text{-}7}$	3 487	3～7	332	100
$T_{8\text{-}12}$	4 732	8～12	197	100
$T_{11\text{-}15}$	4 982	11～15	130	100

3）中文预处理

中文预处理的目的是将测试数据集中的文本转化为结构化的形式，中文预处理包括中文分词、停用词处理和词性标注等步骤。这些预处理可以使用一种中文处理软件来完成。

4）实验方案

实验以中文句子作为基本块，分为两组实验：

（1）实验 1：LDA 模型与多种相似度度量方法及边界估计策略相结合，对文本进行分割，得到实验结果，具体步骤如下：

① 将 LDA 模型运用到 4 个测试数据集上，以句子 s 作为式（6-8）中的文本 d，遍历待分割文本的所有词，运行 Gibbs 抽样算法，迭代足够多次。

② 按照式（6-9）计算主题分布概率 θ 的值。

③ 基于主题分布概率 θ 的值，使用不同的相似度度量方法计算句子间的相似度值 Sim。

④ 结合不同的边界估计策略，通过句子间相似度值 Sim 识别段落的边界。

（2）实验 2：LDA-I 模型运用到 4 个测试数据集上，对文本进行分割，得到实验结果，具体步骤如下：

① 将 LDA-I 模型运用到 4 个测试数据集上，以句子 s 作为式（6-11）中的文本 d，遍历待分割文本的所有词，运行 Gibbs 抽样算法，迭代足够多次。

② 当算法收敛时，段落的边界也同时被标识出来。

对两组实验结果进行比较，评价两种 LDA 模型的性能表现，性能表现采用 P_k 和 WindowDiff（WD）度量来评价。P_k 和 WD 的取值在 0～1 之间，P_k 和 WD 值越小，模型性能越好。为了便于比较和阅读，将每个 P_k 和 WD 值乘以 100。

其中，相似度度量方法包括余弦系数、Jaccard 系数、L1 距离、Hellinger 距离、Clarity 系数和 Jensen-Shannon 发散 6 种方法，边界估计策略包括常数法、动态常数法、局部最小值法和动态规划法 4 种方法。

2．LDA 与 LDA-I 模型性能对比

为了叙述方便，在下面的实验结果中引入了如表 6-2 所示的记号。

表 6-2　记号的含义

记　号	含　义
Cos	余弦系数
Ja	Jaccard 系数
L1	L1 距离
Hel	Hellinger 距离
KL	Clarity 系数
JS	Jensen-Shannon 发散
Con	常数法（常数依次取 0.015，0.020，0.035，2.12，0.1，1.56）
DyCon	动态常数法
Loc	局部最小值法
DP	动态规划法
P_k	P_k 评价指标
WD	WindowDiff 评价指标
LDA(x，y)	LDA 模型，且使用的相似度度量方法为 x，使用的边界识别策略为 y
LDA-I	LDA-I 模型

为了更好地展示实验结果及差异性，实验结果采用两种对比方式：一是 LDA(x，y)在某种边界估计策略下采用不同相似度度量方法与 LDA-I 模型的对比，称为纵向性能对比；二是 LDA(x，y)与 LDA-I 模型在不同相似度度量方法和边界估计策略下的对比，称为横向性能对比。

1）LDA(x，Con)与 LDA-I 性能对比

LDA(x，Con)表示在实验中 LDA 模型采用基于常数法的边界识别策略，并结合不同的相似度度量方法。在实验 1 与实验 2 中，Gibbs 抽样的主题数目 $T = 75$，参数 $\alpha = 50/T$，$\beta = \gamma = 0.01$。取 10 个不同的初始值运行算法，每个初始值迭代 1000 次，然后每隔 100 次取一次样本，共取 10 次样本。每个文本的测试结果取 100 个样本的平均值，实验结果取所有文本测试结果的平均值，见表 6-3。

表 6-3　LDA(x，Con)与 LDA-I 性能对比

模　型	$T_{3\text{-}11}$		$T_{3\text{-}7}$		$T_{8\text{-}12}$		$T_{11\text{-}15}$	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
LDA(Cos，Con)	8.49	19.84	9.20	17.19	6.52	12.37	8.55	19.50
LDA(Ja，Con)	8.42	18.80	8.89	16.35	6.41	12.10	8.40	19.30
LDA(L1，Con)	9.42	17.41	10.90	23.54	11.44	31.24	8.54	15.87
LDA(Hel，Con)	8.39	16.79	10.78	20.37	7.86	25.85	6.71	23.14
LDA(KL，Con)	11.09	29.63	11.57	21.95	8.56	29.00	5.90	15.87
LDA(JS，Con)	9.79	30.02	8.15	15.11	11.16	23.32	13.76	32.57
LDA-I	9.89	18.36	9.59	18.65	8.34	27.56	10.47	22.58

从表 6-3 中的数据可以看出，对于基于常数法的边界识别策略，不同相似度度量方法在不同的测试数据集上的 P_k 值和 WD 值是不同的，说明不同的相似度度量方法适合于不同的测试数据集。总体上，LDA(x，Con)的平均边界识别性能要优于 LDA-I。

2）LDA(x，DyCon)与 LDA-I 性能对比

LDA(x，DyCon)表示在实验中 LDA 模型采用基于动态常数法的边界识别策略，并结合不同的相似度度量方法。在实验 1 与实验 2 中，Gibbs 抽样的主题数目 $T = 80$，参数 $\alpha = 50/T$，$\beta = \gamma = 0.01$。取 10 个不同的初始值运行算法，每个初始值迭代 1 000 次，然后每隔 100 次取一次样本，共取 10 次样本。每个文本的测试结果取 100 个样本的平均值，实验结果取所有文本测试结果的平均值，见表 6-4。

表 6-4　LDA(x，DyCon)与 LDA-I 性能对比

模　型	$T_{3\text{-}11}$		$T_{3\text{-}7}$		$T_{8\text{-}12}$		$T_{11\text{-}15}$	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
LDA(Cos，DyCon)	7.26	25.32	9.99	18.78	10.47	25.28	7.63	30.41
LDA(Ja，DyCon)	7.22	24.68	9.47	18.30	10.27	23.50	7.41	30.14

续表

模　型	$T_{3\text{-}11}$		$T_{3\text{-}7}$		$T_{8\text{-}12}$		$T_{11\text{-}15}$	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
LDA(L1，DyCon)	8.71	16.23	13.19	25.94	13.26	32.84	9.44	26.77
LDA(Hel，DyCon)	14.12	23.05	10.66	31.02	7.72	19.09	14.54	34.05
LDA(KL，DyCon)	9.92	17.46	10.64	38.82	12.38	22.24	16.89	32.57
LDA(JS，DyCon)	10.88	30.69	11.16	22.92	9.40	28.68	10.64	26.32
LDA-I	9.26	28.32	10.01	17.29	9.78	23.13	7.52	25.17

从表 6-4 中的数据可以看出，对于基于动态常数法的边界识别策略，不同相似度度量方法在不同的测试数据集上的 P_k 值和 WD 值是不同的，说明不同的相似度度量方法适合于不同的测试数据集。总体上，LDA-I 的平均边界识别性能要优于 LDA(x，DyCon)。

（3）LDA(x，Loc)与 LDA-I 性能对比

LDA(x，Loc)表示在实验中 LDA 模型采用基于局部最小值法的边界识别策略，并结合不同的相似度度量方法。在实验 1 与实验 2 中，Gibbs 抽样的主题数目 $T = 80$，参数 $\alpha = 50/T$，$\beta = \gamma = 0.01$。取 10 个不同的初始值运行算法，每个初始值迭代 1 000 次，然后每隔 100 次取一次样本，共取 10 次样本。每个文本的测试结果取 100 个样本的平均值，实验结果取所有文本测试结果的平均值，见表 6-5。

表 6-5　LDA(x，Loc)与 LDA-I 性能对比

模　型	$T_{3\text{-}11}$		$T_{3\text{-}7}$		$T_{8\text{-}12}$		$T_{11\text{-}15}$	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
LDA(Cos，Loc)	12.69	21.60	11.57	21.95	8.81	29.62	12.17	26.77
LDA(Ja，Loc)	12.20	21.01	11.28	21.14	8.39	29.14	11.49	26.33
LDA(LI，Loc)	11.50	21.81	11.58	21.95	8.52	28.04	15.81	37.68
LDA(Hel，Loc)	12.74	26.78	10.09	19.73	11.20	20.92	10.35	23.13
LDA(KL，Loc)	8.81	17.44	13.55	28.35	6.98	19.50	11.93	25.64
LDA(JS，Loc)	13.14	22.43	11.57	25.46	10.16	23.32	7.62	21.32
LDA-I	10.27	19.28	11.46	20.79	11.73	29.59	7.57	20.19

从表 6-5 中的数据可以看出，对于基于局部最小值法的边界识别策略，不同相似度度量方法在不同的测试数据集上的 P_k 值和 WD 值是不同的，说明不同的相似度度量方法适合于不同的测试数据集。总体上，LDA-I 的平均边界识别性能要优于 LDA(x，Loc)。

4）LDA(x，DP)与 LDA-I 性能对比

LDA(x，DP)表示在实验中 LDA 模型采用基于动态规划法的边界识别策略，并结合不同的相似度度量方法。在实验 1 与实验 2 中，Gibbs 抽样的主题数目 $T = 80$，参数 $\alpha = 50/T$，$\beta = \gamma = 0.01$。取 10 个不同的初始值运行算法，每个初始值迭代 1000 次，然后每隔 100 次取一次样本，共取 10 次样本。每个文本的测试结果取 100 个样本的平均值，实验结果取所有文本测试结果的平均值，见表 6-6。

表 6-6　LDA(x，DP)与 LDA-I 性能对比

模　　型	$T_{3\text{-}11}$		$T_{3\text{-}7}$		$T_{8\text{-}12}$		$T_{11\text{-}15}$	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
LDA(Cos，DP)	11.26	25.6	10.78	20.36	9.75	25.85	12.19	32.23
LDA(Ja，DP)	10.85	25.44	10.49	20.04	9.39	25.27	11.63	31.51
LDA(L1，DP)	13.06	24.92	16.88	35.07	10.00	21.50	15.32	29.45
LDA(Hel，DP)	8.80	23.68	13.16	25.12	12.64	28.13	6.71	30.41
LDA(KL，DP)	12.73	34.10	10.16	26.92	12.06	27.92	13.99	35.87
LDA(JS，DP)	10.88	21.19	13.16	25.13	11.16	17.92	7.63	30.41
LDA-I	10.93	23.19	10.26	20.25	12.85	22.38	11.59	34.86

从表 6-6 中的数据可以看出，对于基于动态规划法的边界识别策略，不同相似度度量方法在不同的测试数据集上的 P_k 值和 WD 值是不同的，说明不同的相似度度量方法适合于不同的测试数据集。总体上，LDA-I 的平均边界识别性能要优于 LDA(x，DP)。

从上述的实验结果来看，除了 LDA(x，Con)外，LDA-I 模型在其他边界识别策略上的性能表现均优于 LDA 模型。

5）横向性能对比

为了简化起见，在比较不同相似度度量方法和边界识别策略的模型性能表现时，只使用 WD 指标来评价，考察哪种方法或策略的 WD 值最小，其性能表现也就最佳。

首先，考察两种模型采用不同相似度度量方法的 WD 指标表现，如图 6-5 所示，图中的 Cos、Ja、L1、Hel、KL、JS 分别为 LDA 模型所采用的相似度度量方法，New 为 LDA-I 模型的相似度度量方法。由图 6-5 可以看出，在测试集 $T_{3\text{-}11}$ 上，L1 度量的 WD 值最小；在测试集 $T_{3\text{-}7}$ 上，JS 度量的 WD 值最小；在测试集 $T_{8\text{-}12}$ 上，Ja 度量的 WD 值最小；在测试集 $T_{11\text{-}15}$ 上，KL 和 L1 度量的 WD 值最小，而 LDA-I 模型的相似度度量（New）处于中等水平。

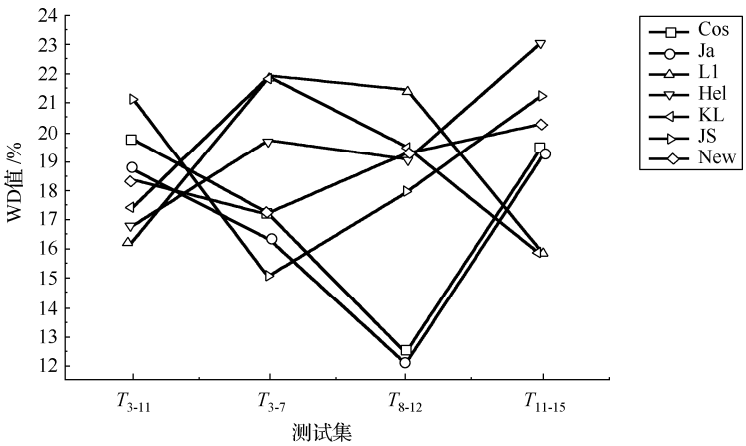


图 6-5　两种模型采用不同相似度度量方法的性能对比

然后，考察两种模型采用不同边界识别策略的 WD 指标表现，如图 6-6 所示，图中的 Con、DyCon、Loc、DP 分别为 LDA 模型所采用的边界识别策略，New 为 LDA-I 模型的边界识别策略。由图 6-6 可以看出，在测试集 $T_{3\text{-}11}$ 上，DyCon 法的 WD 值最小；在测试集 $T_{3\text{-}7}$、$T_{8\text{-}12}$ 和 $T_{11\text{-}15}$ 上，Con 法的 WD 值都是最小的。虽然 Con 法在所有的测试集上的 WD 指标表现最好，但该方法的随机性比较大，不易控制。除了 Con 法外，LDA-I 模型在 $T_{3\text{-}7}$ 和 $T_{11\text{-}15}$ 测试集上的 WD 值最小，边界性能识别性能表现最佳。

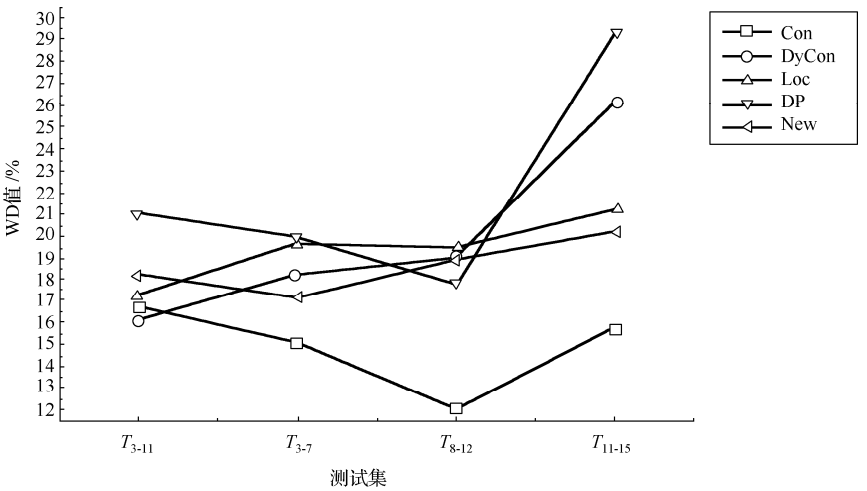


图 6-6　两种模型采用不同边界识别策略的性能对比

从上述的横向性能对比结果来看，LDA 模型的相似度度量方法和边界识别策略在不同的测试数据上的性能表现不同，即它们的性能表现依赖于特定的测试数据集。

由于 LDA-I 模型的相似度度量方法和边界识别策略不依赖于特定的测试数据集，因此在 4 个测试数据集上均有良好的性能表现，并且文本分割的错误率维持在一个很低的水平，说明通过对 LDA 模型的改进，使 LDA 模型的适用性更广，性能更加稳定。

6.4　基于 VSM 模型的文本分割

向量空间模型（VSM）是文本处理中常用的表示模型，也可应用于文本分割中。下面介绍 VSM 模型在文本分割中的应用。

6.4.1　特征项选取

在 VSM 模型中，特征项选取非常重要，直接影响到分类结果的质量。一个理想的特征项集合应该具有两个基本性质：完全性和区分性。

从文本的语法层次来看，一个文本由词、短语、句子和段落等要素组成，所有的要素都可以作为文本的特征。然而，随着这些要素所处的语法层次不断地增高，组合出的特征将会呈指数增长。因此，一般很少采用句子和段落作为特征，常用的文本特征主要有词、短语、*N*-gram 项等。

（1）词语特征。最直观的方法是采用词和短语作为特征来表示文本。在英文中，不需要分词操作，词语已经被空格隔开，但是词语存在词形的变化，如时态变化和单复数变化等，必须进行“词干抽取”处理。对于中文，需要经过分词处理后才能识别出文本中的词语。

（2）*N*-gram 项。在英文中，*N*-gram 项可以由相邻字母组成，也可以由相邻的单词组成。在中文中，*N*-gram 项一般由相邻的字组成，例如，对于“中国人”，如果提取 2-gram 项，则是“中国”和“国人”；如果提取 3-gram 项，就是“中国人”。采用 *N*-gram 项作为特征来表示文本，可以避免复杂的分词处理过程，并且 *N*-gram 项提取比较容易。然而，随着 *N* 的增大，*N*-gram 项的数量将会呈指数增长，大大增加了算法执行时间，影响了效率，因此 *N* 不宜过大。

由于文本中词语数量非常大，采用 VSM 模型表示后，向量空间的维数比较大，这就需要对维数进行降维或者压缩处理。特征向量空间维数压缩方法通常分为特征选择和特征抽取两种。

1．特征选择

特征选择是指排除与类别无关或关联性不大的特征。通常的做法是构造一个评估函数，对特征项集合中的每个特征项进行独立的评估，这样就可以得到每个特征项的评分值（即权值），然后对所有的特征项按照权值大小进行排序，选择预定数量的特征项作为特征选择的结果。因此，选取合适的评估函数和预定特征项数量是特征选择的关键所在。

具体的特征选择步骤如下：

（1）从训练文本集合中提取所有的特征项，构成文本特征集合 *C*；

（2）采用评估函数对集合 *C* 中的每一项进行评分，所有项的评分完成后，按照分值从高到低进行排序；

（3）假设需要选取 *N* 个文本特征项，从 *C* 中依次选取高分值的 *N* 个项，构成最终的特征项集合 C_1，C_1 应用到训练和测试中。

在 VSM 模型中，常用的评估函数有文档频率（DF）、信息增益（IG）、互信息（MI）、卡方检验（CHI）等，详细介绍见 2.4.1 节。

2．特征抽取

特征抽取是指通过对原始特征的合并或转换而得到新的特征。文本分割后的词语经常会出现一词多义、歧义、同义等情况，使得特征集存在很大的冗余，通过特征抽取可以消除

冗余的特征，达到降维的目的。常用的方法有特征聚类和隐含语义标引两种。

（1）特征聚类是将语义相似度高的特征词重新组合成或替换成新的特征词，作为向量空间的特征。特征聚类不同于特征选择，前者关注同义词和近义词问题，而后者关注词对类别区分的信息量大小。

（2）隐含语义标引是基于文本中词与词之间存在的某种潜在语义结构，采用统计方法来寻找语义结构，并用该语义结构来表示词和文本，消除词之间的相关性，较好地解决了词的同义、近义以及多义等问题。

由于词语特征具有丰富的语义信息，因此采用词语作为文本特征具有典型意义，同时可以采用文档频率（DF）等评估函数进行特征降维处理。

基于上述的原理，设文本集合 $D = \{D_1, D_2, \cdots, D_n\}$，所有特征项集合 $T = \{T_1, T_2, \cdots, T_n\}$，这样每个文本可以表示为向量的形式：$D_j = \{W_{1j}, W_{2j}, \cdots, W_{mj}\}$，其中 $W_{ij} = \text{Weight}(T_i)$，表示特征项 T_i 在文本 D_j 中的权重。

6.4.2 语义段分割方法

通过对大量的中文语料文本分析发现，文本的语义结构具有如下特点：大多数文本的主题并不是单一的，一个主题通常又分成若干个子主题，这些子主题从各个方面和不同角度围绕主题进行阐述，子主题可以通过文本中不同的自然段落表达和描述。因此，在段落之间存在语义联系，将语义相同或相近的段落称为语义段，以语义段为分割单元来分割文本，有助于对文本进行细化分析。

1．语义段概念

通常，一篇文章或文本的结构包括以下几个部分：

（1）文本的主题数，即文本由若干个相对独立的主题组成；

（2）文本中的各个子主题或段落所属的主题；

（3）各个子主题或段落之间的相关程度。

按照语义内容的相近程度，可以将文本分成多个语义相对内聚的块，称为语义段，每个语义段由语义上相近的若干个自然段落合并而成，即每个语义段包含一个或多个自然段落。对应于一个子主题，将若干个子主题统一在相关的主题下。

将文本分割成语义段时，需要解决两个关键问题：主题边界的自动识别和分割单元（即语义段）数量的确定。主题边界识别主要涉及相似度度量方法和边界识别策略，详细介绍参见 6.3 节。分割单元数量则需要通过实验来确定。

2．相似度计算方法

经过统计发现，作者在文章中阐述一个主题时，所用的重点词通常局限在一个较小范围内，具有一定的重复性。如果两个段落所含的词语，特别是高频词，发生了一定程度的重

复，则表明段落之间具有较大的相似度，在 VSM 模型中表现为这两个段落的夹角较小，夹角余弦值较大，可以认为这两个段落是围绕同一主题来论述的，应当划分在同一个语义段中。而不同主题的段落所含的词语一般不太相同，在 VSM 模型中表现为这两个段落的夹角较大，夹角余弦值较小，段落之间的相似度较小。

基于这一事实，在 VSM 模型中，可以采用余弦系数法对一个文本中的所有段落进行相似度计算。如果某个段落与前面连续的若干个段落相似度都较小，而与后面连续的若干个段落相似度都较大，则可以认为该段落就是主题划分段。另外，还有一种特殊情况，如果某一段落同时与前面和后面连续若干个段落的相似度都较小，则需要再看该段落的下一段落与其后面的连续若干个段落的相似度是否都较大，如果是，则把该段落也当作主题划分段，因为该段落极有可能是标题段，而标题段含有的词汇少，通常与其他段落的相似度都较小。

假设全文共有 n 个段落，记为 $P_1, P_2, \cdots, P_n$。首先统计词频，去掉低频词和禁用词之后，考察的特征词共有 m 个。将文本中的一个词视为向量空间中的一个维度，段落 P 就可视为 n 维空间中的一个向量 $P(W_1, W_2, \cdots, W_n)$。

3．语义段分割算法

由于相同子主题的段落在语义上必然是相近的，因此可以采用聚集语义相近段落的方法来划分文本中的主题。该方法通过计算整个文本中每两个自然段落之间的语义相似度来判断它们在内容上的相似度，而不再局限于相邻自然段落的相似度比较。如果两个段落之间的语义相似度越大，则越有可能表述的内容是同一子主题。因此，段落之间语义相似度大小的变化可以反映出文本中子主题的变换。

由于语义段是从语义的角度对文本的划分，通过识别各个子主题的边界，文本被自然划分成若干个语义段，每个语义段代表一个子主题，而这些子主题与整个文本表达的主题是相关的。因此，这种基于语义角度的主题划分，有助于对文本进行细粒度的分析。

在介绍语义段分割算法之前，首先给出几个相关的定义如下。

定义 6-1　段落特征向量 $F(P_i) = (W_{i1}, W_{i2}, \cdots, W_{ij}, W_{ik,})$，其中 W_{ij} 表示文本特征词列表中第 j 个元素在段落 i 中的权值，k 为特征向量元素的个数。文本特征向量 $F(D) = (W_1, W_2, \cdots, W_i, W_k)$，其中 W_i 表示文本特征词列表中第 i 个元素在全文中的权值。

定义 6-2　段落与段落之间的语义相似度用下式来表示：

$$M(P_i, P_j) = \frac{\left[F(P_i) \cdot F(P_j)\right]^2}{\left[\left\|F(P_i)\right\| \cdot \left\|F(P_j)\right\|\right]^2} \tag{6-19}$$

$$F(P_i) \cdot F(P_j) = \sum_{m}^{1} W_{im} \times W_{jm} \qquad W_{im} \in F(P_i),\ W_{jm} \in F(P_j) \tag{6-20}$$

$$\|F(P_i)\| = (W_{i1}^2 + W_{i2}^2 + \ldots + W_{ik}^2)^{\frac{1}{2}} \tag{6-21}$$

为了直观地考察文本中各个段落之间的关联情况，以矩阵形式列出各个段落的相似度。由于每个段落与自己的相似度值为 1，因此所得到是一个主对角线为 1 的对称矩阵，其中第 i 行第 j 列的值表示为 $M(P_i, P_j)$，并且满足 $M(P_i, P_j) = M(P_j, P_i)$。

例如，两篇文本的题目分别为“北京奥运安保工作全面展开”和“奥运反恐演习”，它们各自都包含三个自然段落。将这两篇文本合并为一篇长文本，全文共包括 6 个段落。该长文本描述了两个非常相近但又不同的主题，其中前三个段落描述一个主题，后三个段落描述另一个主题。对该长文本中的各个段落按上述算法计算语义相似度，并构建段落语义相似度矩阵。由于该矩阵是对称矩阵，因此在表 6-7 中只列出了矩阵的上三角阵。

表 6-7 中可以看出，P_1 与 P_2 和 P_3 的语义相似度值分别为 0.98 和 0.69；P_4 与 P_5 和 P_6 的语义相似度值分别为 0.75 和 0.67。这说明前三个段落的关联度较大，后三个段落的关联度也较大。P_4 和前三个段落的相似度值分别为 0.11，0.12 和 0.15，相似度都比较小，而它与后面两个段落 P_5 和 P_6 的语义相似度都比较大。因此可以认为段落 P_4 是语义段的分割段。

表 6-7　段落相似度矩阵

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	1	0.98	0.69	0.11	0.10	0.08
P_2		1	0.77	0.12	0.09	0.14
P_3			1	0.15	0.13	0.13
P_4				1	0.75	0.67
P_5					1	0.89
P_6						1

根据上述的分析，基于 VSM 的语义段分割算法（简称 PSBV 算法）如下：

（1）假设待分割的文本中共有 n 个自然段落，记为 $P_1, P_2, \cdots, P_n$。首先对文本进行预处理操作，包括分词、词性标注以及命名实体识别等，获得每个段落中所包含的词条以及相应词频信息。然后统计文本中包含的自然段落数量并做标记，并完成 VSM 模型的构建。

（2）计算文本中每两个段落之间的语义相似度 $M(P_i, P_j)$ $(1<i, j<n)$，得到各个段落的相似度矩阵。

（3）逐个找出主题发生转换的段落候选点 $P_{k_1}, P_{k_2}, \cdots, P_{k_r}, \cdots, P_{k_R}$，其中，$P_{k_r}$ 满足如下的条件：

$$\frac{\sum_{k_{r-1} \leqslant i \leqslant k_r} M(P_i, P_{k_r})}{(k_r - k_{r-1})} < \delta_1 \tag{6-22}$$

$$M(P_{k_r}, P_{k_r-1}) < \delta_2 \tag{6-23}$$

$$k_r - k_{r-1} > 1 \tag{6-24}$$

式中，k_r 是第 r 个语义段分割候选点的段落下标。

式（6-22）表示当前语义段分割候选点与从上一个语义段分割候选点到上一个相邻段的段落相似度平均值要小于阈值 δ_1；式（6-23）式表示当前语义段分割候选点与上一个语义段分割候选点的段落相似度要小于阈值 δ_2；式（6-24）表示当前语义段分割候选点与上一个语义段分割候选点至少要间隔一个段落。

（4）对于语义段分割候选点 P_{k_r}，如果满足下列条件：

$$\frac{\sum_{k_r < i < k_{r+1}} M(P_{k_r}, P_i)}{(k_{r+1} - k_r - 1)} > \delta_3 \tag{6-25}$$

则确定 P_{k_r} 为语义段的分割点，继续处理下一个候选点，直到全部的语义段分割候选点都处理完毕。如果不满足上述条件则转（5）。

式（6-25）表示当前语义段分割点与从下一个相邻段到下一个语义段分割候选点的上一个相邻段的段落相似度平均值大于阈值 δ_3。通过多次实验，分别确定阈值 $\delta_1 = 0.23$，$\delta_2 = 0.31$，$\delta_3 = 0.42$。

（5）如果该语义段分割候选点的下一个段落 P_{k_r+1} 满足如下条件：

$$\frac{\sum_{k_r+1 < i < k_{r+1}} M(P_{k_r+1}, P_i)}{(k_{r+1} - k_r - 2)} > \delta_3 \tag{6-26}$$

则认为 P_{k_r} 为语义段的分割点，否则不是分割点，返回（4）并继续处理下一个语义段分割候选点。

式（6-26）表示当前语义段分割点的下一个段落与从其下一个相邻段到下一个语义段分割候选点的上一个相邻段的段落相似度平均值要大于 δ_3。

通过上述的算法步骤，确定文本中的语义段分割点，将文本中的所有自然段落合并为若干个语义段，这时，文本可以表示为 $\text{text} = S_1 \cup S_2 \cup \cdots \cup S_n$，$S$ 表示语义段。

6.4.3　算法验证

下面通过实验数据对 PSBV 算法性能进行测试和分析。

1．实验数据集

实验数据集是一个关于手机产品评论的中文语料，共有 1 500 篇文本，平均每篇文本的自然段落数为 9.7 个，将所有的文本转换为统一的文本格式，手工将文本分割成语义段并加标注。数据集分为两部分，900 篇作为训练语料集，600 篇作为测试语料集。

为了客观地评价算法的性能，从测试语料库 T 中随机选取 30%、60%的测试语料，形成两个子测试集 T_1 和 T_2，下面的实验分别在 T_1、T_2 和 T 上进行。

2．算法性能对比

在实验中，采用 TextTiling 算法作为实验的对比算法，采用准确率（P）、召回率（R）及 F_1 值评价算法性能。由于没有预先设定分割单元的具体数目，PSBV 算法和 TextTiling 算法需要在识别语义段边界的基础上自动确定文本语义段分割的最佳数目，表 6-8 为实验结果。

表 6-8　语义段分割算法性能对比

测　试　集	PSBV			TextTiling		
	P	*R*	F_1	*P*	*R*	F_1
T_1	0.736	0.723	0.729	0.625	0.501	0.556
T_2	0.747	0.741	0.744	0.619	0.543	0.579
T	0.760	0.759	0.759	0.607	0.577	0.592

从表 6-8 的实验结果可以看出，PSBV 算法的平均召回率达到了 74.1%，平均准确率达到了 74 .8%，平均 F_1 值达到了 74.4%，各项指标都优于 TextTiling 算法。并且随着测试集规模的增大，PSBV 算法的性能也随之提高，在测试集为 T_2 的性能表现比测试集 T_1 好，在测试集为 T 的性能表现比测试集 T_2 好。实验结果表明，采用 PSBV 算法来分割语义段是可行的，并且具有较好的性能。通过将文本分割成语义段，为细粒度地分析文本情感提供了基础。

第7章

文本情感分析技术

7.1 引言

网络文本内容通常是对一些新闻时事、社会热点、法规政策、公众人物、消费产品等话题的个人评价，反映了网民个体的观点。由于每个网民的立场、出发点、个人偏好的不同，对现实世界中各种事物和事件所表达出的立场、态度、意见和情绪的倾向性必然存在很大的差异。对于相同的事物或事件，不同的人从不同视角出发，往往持有不同的观点和态度。例如，对于某个产品，一些用户可能因喜欢该产品某些功能或特点而给出正面（即褒义）评价；而另一些用户也可能因不喜欢该产品某些功能或特点而给出负面（即贬义）评价。

面对互联网中的海量信息，不仅在数量上难以逐一浏览，而且由于受到用户主观认识影响，往往表现出复杂多样的特点。如何快捷而准确地了解人们对某一产品、事件或政策等所持的观点是褒义还是贬义，必须借助于自动化分析技术，这种技术就是文本情感分析技术。

文本情感分析技术主要研究如何对文本所表达的观点、情感、立场、态度等主观性信息进行自动分析，从海量文本中识别出人们对某一产品、事件或政策等所持有的观点是褒义还是贬义，提高对文本情感分析的效率。文本情感分析技术涉及自然语言处理、计算语言学、人工智能、机器学习、信息检索、数据挖掘等多个研究领域，属于交叉性技术。

文本情感分析技术作为一种组织和管理数据的有效手段，在网络舆情监测、电子商务等领域中得到了应用。随着互联网发展和大数据时代的到来，将会不断地扩展文本情感分析技术的应用领域。

在网络舆情监测中，对于一个突发社会公共事件的网络舆情，网民所持有的态度倾向性往往是多元化的，包括正面或负面、赞扬或批评、支持或质疑、肯定或否定等。通过文本情感分析技术，能够自动识别出其态度倾向性，并给出分类统计结果，有助于及时采取应对措施。

在电子商务中，通过文本情感分析技术，对产品评论的褒贬倾向性进行分析，可以帮

助生产者和商家及时了解消费者对产品的反馈意见，从而做出准确的商业决策。例如，在电子商务推荐系统中，通过分析消费者对产品及服务的反馈评论和意见，向其他用户推荐受到好评的产品和服务，同时利用用户的反馈信息，对消费市场进行深入分析，对产品和服务进行总结和改进。

文本情感分析技术主要研究如下 3 个问题：

（1）主客观分析，即识别出文本内容是主观性表达还是客观性表达，文本情感分析主要关注主观性表达的文本内容；

（2）情感倾向分析，对于主观性表达的文本内容，识别出情感倾向性：褒义、贬义还是中性；

（3）情感强度分析，即判定文本情感倾向的强弱程度，例如，将情感强度分为强烈贬抑、一般贬抑、客观、一般褒扬、强烈褒扬等类别，判定当前文本的情感倾向所属于的类别。

情感强度分析比较复杂，并且情感强度分类也缺乏被广泛认可的标准。因此，文本情感分析技术主要研究前面两个问题。

需要指出的是，文本情感分析技术主要将文本的情感倾向分为褒义和贬义两类，对于网络舆情分析中来说，还不够细致。在此基础上，还需要人工做进一步的统计分析。

本章将从句子情感分析方法、段落情感分析方法以及文本情感分析模型 3 个层面来介绍文本情感分析技术。

7.2 基本概念

7.2.1 文本情感分析层次

文本情感分析大致可以分为词语情感分析、句子情感分析、文档情感分析等三个层次。

1. 词语情感分析

词语情感分析的对象是在特定的句子中出现的词和短语。表达情感的词大多是名词、动词、副词和形容词，其情感倾向可以分为褒义、贬义和中性等三类，词语情感分析包括对词的情感极性、情感强度以及上下文模式等进行分析。

在词语情感分析时，需要借助于标注有倾向性的情感词典。不论是英语还是汉语，词汇都是非常丰富和庞大的，并且有相当数量的词语随着语境的变换，其倾向性也会发生相应的变化。因此，构建一个涵盖所有情感词倾向性的完整词典是非常困难的，一般都是面向领域应用来构建情感词典。在构建情感词典时，大多采用在已有的电子词典或词库上进行扩展的方式。例如，在知网（HowNet）的词库上进行扩展。

词语情感分析是文本情感分析的基础，主要涉及以下方面的工作：

（1）情感词典的构建及扩展；

（2）语料库的情感倾向标注；

（3）新词语的情感倾向判断；

（4）文本中情感词重要性分析。

2．句子情感分析

句子情感分析的对象是在特定的上下文中出现的句子，其目的是通过分析句子中的各种主观性信息，判断该句子是主观句还是客观句，对于主观句，进一步提取出句子中的主观关系，实现对句子的情感倾向的判断，同时还要分析与情感倾向性相关的各个要素，如评价对象、情感极性、情感强度等。图 7-1 为句子情感分析流程。

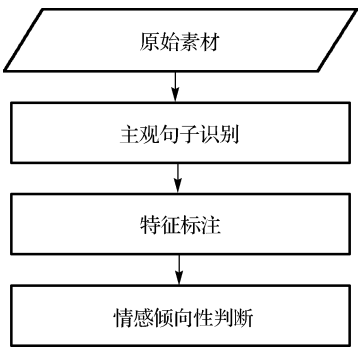


图 7-1　句子情感分析流程

句子情感分析是文本情感分析的重点，主要涉及以下方面的工作：

（1）主题句和主观句识别和提取；

（2）主观句中的主观关系识别和提取；

（3）主观句的情感倾向性判断；

（4）情感倾向性相关要素分析。

3．文档情感分析

文档情感分析的对象是一篇完整的文章，从整体上分析某个文章的情感倾向性。由于文档情感分析属于文本分类问题，通常采用机器学习方法，如朴素贝叶斯、最大熵、支持向量机等方法来解决文本情感分析问题，首先构建语料库，人工标注语料库中每个文本的情感倾向，并将语料库分为训练集和测试集，然后对模型进行训练和算法测试，对模型和算法的文档情感倾向识别能力进行评价。

文档情感分析主要涉及以下方面的工作：

（1）语料库的构建和情感倾向标注；

（2）分类模型选取和模型参数确定；
（3）分类模型训练和测试；
（4）文档全局情感倾向识别。

7.2.2 文本情感分析方法

文本情感分析属于文本分类问题，需要事先建立一个语料库，对语料库中每个文本的情感倾向进行分类标注，通常分为褒义和贬义两大类，然后构造一个分类器，将待分类或分析的文本归入不同的类别中，从而实现对文本情感倾向的分析。

在文本情感分析中，文本分类主要采用自动分类方法，自动分类方法又可分为知识工程方法和机器学习方法两大类。

1．知识工程方法

从概念上，知识工程方法是指由专家为每个类别定义一些规则，这些规则代表了这个类别的特征，自动把符合规则的文档划分到相应的类别中。

对于文本情感分析，事先构建一个面向领域的情感词典，包含该领域中所有可能的评价对象及其特征，通过匹配规则，识别出待分析文本的情感倾向。这种方法的基本思想是，给定一组已知极性的词语集合作为种子，对于一个情感倾向未知的新词，在电子词典中找到与该词语义相近并且在种子集合中出现的若干个词，根据这几个种子词的极性，对未知词的情感倾向进行推断。对于情感词典的构建，一般采用在现有的电子词典或词库上扩充而成。例如，中文的情感词典大多在 HowNet 词典上扩充，英文的情感词典大多在 WordNet、General Inquirer 等词典上扩充。

知识工程方法具有获取情感词全面、准确的优点，但对已有词典的依赖性较强，受种子词数量的影响也比较明显。同时，由于存在一词多义的情况，构建的情感词典中往往包含了较多的歧义词。

2．机器学习方法

在文本情感分析中，主要采用有监督的机器学习算法来识别文本中的评价对象及情感倾向。这种方法需要事先由人工标注语料库的情感倾向，作为训练样本，不同领域的训练样本也不同。然后构造一个分类器算法，经过自动训练后，对待分析文本的情感倾向进行分类识别。这种方法的优点是简单易行、精度较高，在文本情感分析中能够取得较好的效果。但是该方法依赖于人工标注的语料库，而人工标注语料库费时、费力，并且由于缺乏标注标准，语料库标注格式也不统一。

7.2.3 语言建模方法

文本及其情感是通过语言来表达的，语言是随着文化变迁自然演化而成的。在互联网

时代，独特的网络文化衍生了网络语言表达的习惯和特点，使网络文本与普通文本有所不同，具有自身的语言特点和特殊的表达方式。因此，根据网络文本语言自身的特点、表达习惯和特征，可以采用语言建模方法来识别文本情感倾向。

语言建模方法是一种自然语言处理中的重要方法，语言模型大致可以分为如下两类：

（1）基于语法和语义信息的建模方法。这种方法又分为两种：一种方法是利用浅层语法、语义结构信息来建立语言模型，通常只引入较少的语法、语义知识，并且不定义显性的语法结构。例如，基于词类的语言模型、基于 Trigger 的语言模型、潜在语义分析模型、Skipping 语言模型等。另一种是基于语法或者语义结构的语言模型，主要从分析句子语法、语义结构入手来构建语言模型，更多地利用了语言的结构信息。例如，结构语言模型、基于自上而下句法分析器的语言模型、无监督学习的依存结构模型等。

（2）统计语言建模（Statistical Language Modeling，SLM）技术。主要采用统计学和概率论方法对自然语言进行建模，发掘出自然语言中的规律和特性，解决自然语言信息处理中的特定问题。SLM 技术已被广泛应用于语音识别、光学字符识别、手写字识别、机器翻译、文本分类与过滤以及文本检索等诸多领域，成为自然语言信息处理的主流技术之一。

7.3　句子情感分析方法

一个文本是由不同性质和类型的句子组成的，不同类型的句子在文本内容表达中所起的作用也不同，主题句包含了文本主题概念，对文本内容的表达起到重要的作用。从情感分析的角度，可以将句子分为主观句和客观句两大类，主观句是主观性表达，带有作者个人的情感和意向的抒发，反映作者的立场和态度，具有一定的感情色彩。客观句是对事实的客观性陈述，不带有个人的好恶和偏见，具有客观性和确定性。因此，文本情感分析的对象是主观句。在句子情感分析中，主要研究的是主题句、主观句以及主观关系等识别和提取问题。

下面主要介绍主题句、主观句以及主观关系的识别方法，为句子情感分析奠定基础。

7.3.1　主题句识别方法

文本的主题句是指包含文本主题概念的句子，与文本中的普通句子不同，它既是文本中心思想的重要载体，同时也是文本内容的集中体现，对文本内容的表达起到更为重要的作用。

根据其主客观特性，文本中的主题句可以分为两类：一类是包含情感描述并具有褒贬倾向性的句子，称为主题情感句；另一类是不包含情感描述的句子，称为客观主题句。主题情感句不仅包含了文本所要表达的主题，还具有情感倾向性，对于识别句子情感倾向性具有重要的作用。

要识别文本中的主题句，首先需要识别出文本主题。一般来说，语句中的主题有两种

形式：一种是显式主题，可以直接从语句中获得；另一种是隐式主题，需要通过对当前语句上下文的语义分析才能获得。显然，后者的识别具有相当的难度。因此，大多数的主题识别方法主要是面向显式主题的识别。

显式主题主要是领域相关的术语，通常采用了两种识别策略：一种策略是根据短语结构的特点来识别，但这种方法存在主题术语覆盖面的问题。另一种策略是根据候选主题的共现特征和上下文指示符来识别常现和非常现的主题术语。

在主题句识别中，首先需要定义主题概念评估指标，通过这些指标对文本包含的语义概念进行评估，确定该文本的主题概念。然后将所有包含主题概念的句子作为候选主题句，通过计算各个候选主题句的权值，从候选句子中确定文本的主题句，构成文本的主题句集合。

1．主题概念定义与评估

文本的内容是通过一定的语义概念集合来表达的，语义概念是指文本中的基本语义单元，可以是文本中的一个词语，也可以是文本中多个语义相近的词语。通常，文本的语义概念包含两类：主题概念和非主题概念。主题概念是与文本表达的内容和主题思想相关的概念。

文本中的主题句是指包含主题概念的句子。要识别文本中的主题句，首先要确定文本的主题概念，即从文本的语义概念中提取主题概念。

首先，将 HowNet 作为语义概念获取的资源，通过对 HowNet 知识库中的信息进行处理，可以得到 HowNet 中的一些有用信息，从而获取语义概念。例如，对 HowNet 处理后，获得的部分有用信息如表 7-1 所示。

表 7-1　语义处理信息样例

W_X	G_X	DEF
工作	N	fact｜事情，do｜做
工作	N	affairs｜事务，undertake｜担任
饭碗	N	affairs｜事务，＃职位，earn｜赚，alive｜活着
职业	N	affairs｜事务，＃职位，earn｜赚，alive｜活着
差事	N	affairs｜事务，＃职位，earn｜赚，alive｜活着

在表 7-1 中，W_X 表示词语项，词语项中的各个词语就是语义相近的词语，G_X 表示词性，DEF 为词语的定义。由于 HowNet 对其知识库中的所有词语都给出了一个定义，即 HowNet 中的 DEF 项，因此可以定义 DEF 项为该词语的概念。利用 HowNet 知识库中的资源，可以得到每个词语的语义概念。

作为文本的主题概念，首先与文本的内容有一定的联系，具有一定的语义归纳能力，但是又不能过于笼统。为了从语义概念中提取主题概念，需要定义两个评估参数：语义概

念重要度和分布广度，综合考察语义概念的重要度以及归纳能力等因素。然后根据两个评估参数的计算结果，得到语义概念选取度，最终确定一个语义概念是否能够成为文本的主题概念。

1）语义概念重要度

如果同一个概念在文本中出现的次数相对比较多，则说明它是作者在文本中反复提到的描述对象，因而很有可能就是文本的主题或者与文本主题具有较强的相关性。因此，可以使用语义概念 C 在文本中出现的频率来定义一个语义概念的重要度。

设文本中的语义概念 C 的词语集合为 $\{w_1, w_2, \cdots, w_n\}$，$C$ 的重要度计算公式如下：

$$F(C)=\sum_{i=1}^{n} f(w_i) \tag{7-1}$$

式中，$f(w_i)$是词语 w_i 在文本中出现的频率，这个指标反映了表达同一个语义概念的词语出现次数。

2）语义概念分布广度

一个语义概念在文本中分布越广，越有可能与主题有关，因为主题概念通常是贯穿性出现在整个文本中的，与其他非主题概念相比，主题概念在文本中的分布相对比较广泛。因此，将语义概念 C 的词语集合中各个词语在文本段落中的分布疏密程度称为语义概念的分布广度。

设文本的段落数目为 N，则语义概念 C 的分布广度计算公式如下：

$$S(C)=\frac{1}{n}\sum_{i=1}^{n}\frac{d(w_i)}{N} \tag{7-2}$$

式中，$d(w_i)$是文本中含有词语 w_i 的段落数量，这个指标用于综合衡量该语义概念的概括能力。

3）语义概念选取度

为了确定是否选取该语义概念作为主题概念，使用语义概念选取度来评估一个语义概念成为主题概念的可能性。语义概念选取度计算公式如下：

$$\text{Select}(C)=\alpha \lg F(C)+\beta \lg S(C) \tag{7-3}$$

式中，$F(C)$和 $S(C)$分别是概念 C 的重要度和分布广度，而 α 和 β 为加权系数，其作用是调整各个参数之间的权重，α 和 β 的取值需要根据经验并结合实验来确定，例如可以选取 $\alpha = 1$、$\beta = 0.26$。Select (C)值越大，语义概念 C 越有可能是文本的主题概念。

在完成语义概念选取度计算后，需要设置一个阈值，当 Select(C)大于该阈值时，则认为语义概念 C 为主题概念，并将其归入主题概念集合 T 中。该阈值的取值也要根据经验并结合实验来确定，例如可以选取该阈值为 1。

2．句子重要度计算方法

确定文本的主题概念后，将文本中所有包含主题概念的句子都作为候选主题句子归入候选集合中。为了从候选集合中最终确定文本的主题句子，需要对各个候选句子的重要程度进行评估计算，从中选取重要度相对较高的句子作为最终的主题句。

对每一个待处理的候选主题句子 S，将句子中包含的每个词语归入到对应的主题概念上，建立起对应向量 $S\ (T_1, W_1; T_2, W_2; \cdots, T_n, W_n)$，其中 T_i 为句子所含的各个主题概念，W_i 为 T_i 对应的频度。建立空间向量模型，对各个句子的重要度进行计算，句子 S 的重要度计算公式如下：

$$I(S)=\alpha\frac{\sum_{i=1}^{n}w_i}{n}+\beta W_p+\lambda W_c \tag{7-4}$$

式中，W_p 为句子的位置加权系数。α、β 和 λ 是线性加权因子，且满足 $\alpha+\beta+\lambda=1$。W_c 表示提示词的影响系数。

句子的位置是决定句子重要性的一个重要因素，不同位置的句子加权系数不同。就中文而言，由于中文文本表达通常讲究“起、承、转、合”，因此出现在段首和段尾的句子通常比出现在其他位置的句子更为重要，这些句子往往概括了整个文章或一个自然段落的中心内容，表达或强调作者的观点，比其他位置的句子的重要度高。根据中文文本表达的特点，将下列 4 种类型的句子赋予较高的权值，以表明这 4 种类型的句子具有较高的重要度。

（1）标题句，即在句子中含有在标题中出现的有效词。

（2）含有高频有效词的句子。

（3）文中重要位置的句子，如段首句、段尾句以及第一个自然段中的句子等。统计显示，段落的论题是段落首句的概率为 85%，是段落尾句的概率为 70%。

（4）含有提示短语的句子。对于一些议论或评述性的文本，常常包含提示性词或者短语，例如，含有“总之”、“因此”、“所以”、“论述了”、“表达了”等短语，包含这些提示词的句子往往是对文本的主题进行概述，其重要度相对较高。

对于式（7-4）中各个参数值，需要根据句子的位置等因素，并通过实验来确定。对于标题句，W_p 取值为 1；对于段首和段尾的句子，W_p 分别取值为 1 和 0.5。W_c 表示提示词的影响，如果该句包含了提示词，则 W_c 取值为 1，否则取值为 0。

在各个参数取值确定后，就可以用式（7-4）来计算候选集合中所有句子的重要度，然后按照重要度从大到小对句子进行排序。

3．主题句去重处理

主题句的选择是根据句子的重要度大小来确定的，通常选择重要度相对较大的一些句子作为文本的主题句。在文本中，作者为了强调文章主题的内容，往往会使用一些重复性表

述的句子，而这些句子往往都具有较大的权值。根据实验显示，在抽取主题句时难免会出现抽取出多个反映相同主题的主题句，即主题句冗余问题。因此，对于重复的主题句，需要进行去重处理。

重复主题句是指那些都表达了文本主题但相似度较高的句子。在决定句子的取舍时，需要考虑文本包含的主题数量、所抽取的主题句数量等因素。首先需要满足主题的覆盖率，即每个主要的主题都有一个对应的主题句。在此基础上，如果抽取的主题句较多，则需要舍弃那些主题相似度较高的句子。

在主题句去重处理时，首先需要确定所抽取的主题句数量，然后采用余弦系数法来计算候选主题句的相似度。文本主题句的抽取数量与具体的应用有关，例如，对于文本标题自动生成，通常只需要抽取出文本中最具有代表性的一个主题句；对于文本自动摘要生成，抽取句子的数量通常是原始文本句子总数的 20%～30%；对于文本情感分析，既要保证抽取的主题句具有一定的主题覆盖率，还要求抽取句子的数量不能过少，以免影响到分析的准确率，抽取主题句的数量应不低于文本中句子总数的 45%。

4．主题句选择算法

对于主题句选择算法，主要考虑抽取句子的主题覆盖率和概括性，使抽取的句子能够反映文本的主要内容，并使发生主题遗漏和内容重复的可能性最小。主题句选择算法的主要步骤如下：

（1）首先采用余弦系数法计算候选主题句之间的相似度，并预先设定抽取数量 N；

（2）对于具有一定相似度的一组句子，仅选择重要度最大的一个句子作为主题句，并从候选主题句集合中删除该句，但不删除其余的句子；

（3）对于相似度非常高的一组句子，仅选择重要度最大的一个句子作为主题句，其余的句子都从候选主题句集合中删除。

（4）如果抽取数量未达到 N，则按句子重要度逆序从候选主题句集合中选择一个或多个主题句。

按照上述的步骤，提取出文本中的主题句，并将所有提取的主题句归入主题句集合 Q 中。

7.3.2　主观句识别方法

不论主题句还是非主题句，都存在主观性表达和客观性表达两种表达方式。客观性表达是对事实的客观性陈述，不带有个人的好恶和偏见，具有客观性、绝对性和确定性。主观性表达则基于断言或评论，一般带有作者个人的情感和意向的抒发，反映作者的立场和态度，具有一定的感情色彩。主观性表达可以分为两类，一类是评价，主要包括希望、仇恨、喜欢、讨厌等多种感情，以及评论、判断和意见等；另一类是推测，主要是指非现实中发生的或非实际持有的心理状态。通常所说的主观性表达大多属于第一类。

由于文本中的客观性表达是对客观事实的陈述，不带有任何感情色彩；而主观性表达则反映了作者的褒贬倾向性。因此，文本情感分析的对象主要是主观性表达的句子，称为主观句。

1．主观句与主观关系

首先需要解决的问题是如何识别文本中的句子是客观性表达还是主观性表达，即客观句还是主观句。在主观性表达中，能够体现主观性的匹配关系称为主观关系。主观关系通常由两部分组成：评价对象和评价词语。评价对象一般为名词或名词词组，而评价词语一般为形容词、副词以及少量动词。评价词语的作用是对评价对象的描述和修饰。例如，在“苹果手机外形漂亮”的句子中，苹果手机外形是评价对象，评价词语为形容词“漂亮”，在“漂亮”和“苹果手机外形”之间存在一种主观关系，体现了作者对评价对象的态度和立场。

由于主观关系只存在于主观性表达中，而主观性表达在文本中通常是以主观句的形式出现的，因此需要首先识别出文本中的主观句，然后才能挖掘出主观句中存在的主观关系，进而根据主观关系对文本情感进行分析。

2．情感词库构建方法

通常，一个主观句至少包含一个情感词，可以通过判断一个句子中是否含有情感词来确定该句子是否属于主观句。因此，主观句识别的关键在于句子中的情感词识别。

情感词是指带有情感色彩和褒贬倾向性的词，情感词具有以下特点：一是情感词语主要是形容词、副词以及少量的动词等；二是在同一个单句中出现的情感词通常具有相同的褒贬倾向性，一般由“、”或者“和”连接，这就是词的共现特性。

对于情感词识别，首先需要建立一个情感词的词典或者词库。由于词汇的丰富性，建立一个包含所有情感词的词典是不现实的。应当根据情感词自身的特征，建立一个面向领域或应用的情感词库。

首先，确定具有鲜明褒义和贬义的种子词语，以种子词语为基础，扩展情感词，对扩展后的词语进行评价选择，获得新的情感词语，从而扩展构建面向领域的情感词库。

情感词库的构建方法如下：

（1）基本情感词库建立。利用 HowNet 建立基本的情感词库，HowNet 包含有 6 564 个词条，需要为每个词条添加两个属性：褒贬倾向和情感值。褒贬倾向包含：Positive（褒义）、Negative（贬义）和 Neutral（中立），由人工完成褒贬倾向和情感值的标注，Positive 的标注值为 1，Negative 的标注值为–1，而 Neutral 不标注。这样就保证了绝大多数常用的情感词都可以在 HowNet 中直接检索到。

HowNet 是一个以汉语和英语的词汇所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。对于汉语词汇，HowNet 中的描述是基于“义原”概念的。义原是指汉语中不能再分割的最小语义单位。因此，汉语中的

词可以理解为若干义项的集合。在 HowNet 的语义字典中，每条记录都是由一个词的一条义项及其描述所组成，即一条记录对应于一个词的一个义项。

（2）情感词选取选。对语料库中的文本进行分词，从中选出所有的情感词，共选出 1 255 个情感词。手工进行褒贬倾向分类，从中选取分类一致的结果，同时忽略有歧义的词语，从而得到 623 个 Positive 词，情感值标注为 1；另外得到 354 个 Negative 词, 情感值标注为–1。这样就得到褒义种子集合和贬义种子集合。

（3）情感词库扩充。在实际应用中，文本中往往会出现情感词库中不存在的情感词，称为新情感词。通过识别新情感词的褒贬倾向并标注后，添加到情感词库相应的褒义种子集合和贬义种子集合中，实现对情感词库的扩充。

3．新情感词识别方法

对于新情感词的褒贬倾向值，可以利用下式来计算：

$$\mathrm{SO\text{-}PMI}(w)=\sum_{W\in \mathrm{PSET}}\mathrm{PMI}(w,W)-\sum_{W\in \mathrm{NEST}}\mathrm{PMI}(w,W) \tag{7-5}$$

式中，PMI（Pointwise Mutual Information）表示互信息概率，PSET 和 NSET 分别代表褒义种子集合和贬义种子集合，通过计算一个新词与两个种子词之间的互信息概率，就可以得到该词的褒贬倾向值。其中，PMI 计算公式如下：

$$\mathrm{PMI}(w_1,w_2)=\log\frac{p(w_1\bigcup w_2)}{p(w_1)\times p(w_2)} \tag{7-6}$$

式中，$P(w_1\bigcup w_2)$表示 w_1 和 w_2 同时出现的概率。

如果新词与其前面或后面的情感词之间出现了关联词。首先判断关联词的类型，然后根据以下规则确定该词的褒贬倾向：

（1）如果是递进关联词或者是并列关联词，则该词的褒贬倾向与句子中出现的情感词相同；

（2）如果是转折词，则该词的褒贬倾向与句子中出现的情感词相反。

4．主观句识别方法

在情感词库构建完成并确定新情感词识别算法后，可以按照下列步骤完成主观句识别。

（1）对于一个文本中的句子，首先判断句子中是否存在情感词，并提取情感词。

（2）查询情感词库，如果从两个种子集合中找到相应的情感词，则直接判定为该句子为褒义或贬义的主观句。如果找不到相应的情感词，则通过情感词识别算法来判断是否为新情感词；如果是新情感词，则判定该句子为主观句，并将该情感词添加到情感词库相应的褒义种子集合或贬义种子集合中。

7.3.3 主观关系识别方法

在识别出文本中的主观句后，需要识别出主观句中评价词与评价对象之间的主观关系，进而实现对文本情感的分析。在句子主观关系识别中，采用最大熵模型。

1．最大熵模型

最大熵模型的基本思想是，在只掌握未知分布的部分知识时，应选取符合这些知识且熵值最大的概率分布。由于熵的实质是随机变量的不确定性，当熵最大时，说明随机变量最不确定。因此，最大熵的实质就是在已知部分知识的前提下，关于未知分布最合理的推断就是符合已知知识最不确定或最随机的推断，这是一种不偏不倚的选择，而任何其他的选择都会增加一定的约束条件和假设。

设随机过程所有的输出值构成一个有限集，对于每个输出结果 $m \in M$，其产生均受到上下文 c 的影响，c 属于有限集 C。最大熵模型的目标就是对于给定上下文 c，计算出 m 的条件概率，即对 $p(m|c)$进行评估，期望能够求出符合 c 条件的 m 的概率分布。最大熵模型要求 $p(m|c)$在满足一定的约束条件下，必须使下式定义的熵取得最大值：

$$H(p) = -\sum_{c,m} p(m \mid c) \log p(m \mid c) \tag{7-7}$$

最大熵的条件概率可以用下式计算：

$$p(m \mid c) = \frac{1}{Z(c)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(c,m)\right) \tag{7-8}$$

式中，$Z(c)$可以用下式来表示：

$$Z(c) = \sum_{m} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(c,m)\right) \tag{7-9}$$

式中，f_i 是模型的特征，λ_i 是 f_i 的参数，它表明了特征 f_i 对于模型的重要程度，即每个特征函数的权值。特征 f_i 是一个二值函数，用于描述某一个特定的事实，每个特征包含了上下文的各种信息。

在最大熵模型中，每个特征一般由两个部分组成：一部分称为条件或上下文 x，另一部分称为目标概念类，即分类标记 y。实际上，可以描述为“如果…则…”两部分，前者是“条件”描述，后一部分则是目标的描述。特征值一般定义为 $f(x, y)$，如果 x 和 y 满足某种条件，那么 $f(x, y) = 1$；否则 $f(x, y) = 0$。

在最大熵模型构建过程中，需要解决的两个主要问题是约束的确定和参数 λ_i 的求解。约束问题与特定的应用密切相关，而参数 λ_i 值则不能直接得到，需要通过迭代算法计算其近似值。迭代算法可以概括为以下几个步骤：

（1）假定迭代初始的模型为等概率的均匀分布。

（2）用当前迭代得到的模型来估计每种信息特征在训练数据中的分布，与训练数据中的实验分布进行比较，如果超出了实验分布值，则把相应的模型参数值减小；否则，将相应的模型参数值增大。

（3）重复步骤（2），直到模型参数收敛为止。

目前，常用的迭代算法有 GIS（Generalized Iterative Scaling）算法、IIS（Improved Iterative Scaling）算法等。在下面的实验中，使用 GIS 算法来实现迭代，迭代次数为 100。

2．模型特征选取

最大熵模型应用的关键在于如何针对特定的任务为模型选取合适的特征集合。在实际的应用中，采用简单的特征表达复杂的语言现象，承认已知的可观察到的事实，不做任何独立性的假设，将这些观察到的事实表示为最大熵模型的特征集合。

模型特征的选取需要使用特征选择算法。假定所有特征的集合是 F，特征选择算法要从中选择一个活动特征集合 S，活动特征集合要尽可能地准确反映样本信息，只包括那些期望可以准确估计的特征。为得到活动特征集合 S，通常采用一个逐步增加特征的方法，每次要增加哪个特征取决于样本数据。例如，当前的特征集合是 S，满足这些特征的模型为 $C(S)$，增加一个特征后，新的模型集合可以定义为 $C=(S\cup f)$。在特征选择过程中，活动集合越来越小，同时模型集合会越来越大。

在句子主观关系提取中，需要根据主观句自身的特点进行模型特征的选取。主观句中评价词与评价对象的词特征反映了评价词与评价对象的匹配关系，不同的评价对象与评价词之间的修饰关系有一定规律可循。例如，“身材”通常与“高”、“矮”、“胖”、“瘦”之类的评价词搭配，而不会与“便宜”、“昂贵”之类的评价词搭配。因此，对于一个主观句，首先提取该句中所有的评价词，形成一个评价词集合$\{E_1，E_2，\cdots，E_n\}$，然后提取该句子中所有的评价对象，形成一个评价对象集合$\{O_1，O_2，\cdots，O_N\}$。对于评价词集合中的每一个 E_i，根据特征函数 $F_k(O,\{O_1，O_2，\cdots，O_N\}，E_i)$，计算其条件概率 $p(O\mid\{O_1，O_2，\cdots，O_N\}，E_i)$。

与评价词 E_i 之间具有对应匹配关系的评价对象 O 可以通过下式计算得到：

$$O=\operatorname{argmax}\left[\sum_{i=1}^{n}\lambda_i F_i(O,\{O_1,O_2,\cdots,O_i\},E)\right] \tag{7-10}$$

实际上，主观句中主观关系的提取可以看作对句子中的评价词进行主观关系标注的过程。这个标注过程被看作一个事件，由当前评价词及它的上下文语境来确定这一事件的特征集合。

根据当前评价词主观关系标注的各种影响因素，将特征空间定义如下：

（1）词：当前评价词的前后各两个词；

（2）词性：当前评价词及其前后各两个词的词性；

（3）距离：当前评价词与评价对象之间的距离；

（4）语义：当前评价词以及其前后各两个词的语法语义信息。

根据这个特征空间，在模型训练中应用的特征可以定义为两大类：第一类是基本特征，主要描述词本身的特性，具体包括词特征、词性特征以及距离特征等；第二类为语义特征，主要包含句子中的相关语义信息。表 7-2 到表 7-5 分别给出了词特征、词性特征、距离特征以及语义特征。

表 7-2　词特征

特征名称	特征具体描述
WE	评价词
WE1	评价词左前第一个词
WE2	评价词左前第二个词
WE-1	评价词右后第一个词
WE-2	评价词右后第二个词
WO	评价对象
WO1	评价对象左前第一个词
WO2	评价对象左前第二个词
WO-1	评价对象右后第一个词
WO-2	评价对象右后第二个词

表 7-3　词性特征

特征名称	特征具体描述
WE	评价词性
WE1	评价词左前第一个词性
WE2	评价词左前第二个词性
WE-1	评价词右后第一个词性
WE-2	评价词右后第二个词性
WO	评价对象词性
WO1	评价对象左前第一个词性
WO2	评价对象左前第二个词性
WO-1	评价对象右后第一个词性
WO-2	评价对象右后第二个词性

表 7-4　距离特征

特征名称	特征具体描述
P(E-O)	评价对象和评价词语的前后顺序关系
N1	评价词语和评价对象之间间隔的评价对象的个数
N2	评价词语和评价对象之间间隔的评价词的个数
D(E-O)	评价对象和评价词语间隔的词的个数

表 7-5　语义特征

特征名称	特征具体描述
Path(E-O)	节点 Head(E-O)到其左右孩子节点的路径（选取的孩子节点同时是 E 和 O 的父节点）
Path(E)	节点 E 到其任何一个祖先节点的路径（选取的祖先节点同时是 Head(E-O)的孩子节点）
Path(O)	节点 O 到其任何一个祖先节点的路径（选取的祖先节点同时是 Head(E-O)的孩子节点）

词特征和词性特征不仅考虑到评价词和评价对象本身以及它们的词性，并且还考虑了一个词的前后两个邻词，因为前后邻词在一定程度上体现了该词是否具有主观含义。同时，还解决了否定词及程度副词对评价词的影响问题，因为否定词和程度副词通常起修饰作用，一般位于评价词前后两个词范围的位置上。否定词可以改变评价词的倾向性，对主观关系产生一定的影响。程度副词对句子的语义强度产生很大的影响，将程度副词作为特征考虑，既可以在一定程度上反映评价词主观性的强弱程度，而且还可以缩短评价词和评价对象之间的距离，距离越短越容易做出正确的判断，而距离越长越容易发生误判。

距离特征主要描述了评价词和评价对象在一个主观句中所处的位置关系，距离越近，两者之间越有可能存在主观匹配关系。

语义特征主要描述了评价词在句子中的句法语义信息，这是考虑到上下文语境的影响，也是对基本特征的有效补充。语义特征提取过程比较复杂，需要使用中文句法分析器对句子进行句法分析，获得该句子的句法结构树，从中提取评价对象和评价词之间的句法路径信息作为特征，这种路径信息描述了评价词和评价对象在句子语法结构中的位置以及修饰关系，有助于正确判断句子中存在的主观关系。

7.3.4　算法验证

下面通过实验数据对主题句识别算法和句子主观关系识别算法的性能进行测试和验证。

1．主题句识别算法

1）实验数据集

实验数据集是一个关于手机产品评论的中文语料，共有 600 篇文本，手工提取语料中的主题句，并标注各个主题句的褒贬倾向（褒义或贬义）。数据集分为两部分，350 篇作为训练语料集，250 篇作为测试语料集。

实验分为算法性能测试和算法性能对比，算法性能评价指标采用准确率（P）、召回率（R）和 F_1 值。

2）算法性能验证

下面的实验是使用 7.3.1 节中给出的主题句识别算法（简称本方法）对主题句识别效果进行测试和验证。

为了准确地评价算法的性能，从测试语料库 T 中随机选取 20%、50%的测试语料，形成两个子测试集 T_1 和 T_2，各个测试集手工标注主题和褒贬倾向情况统计如表 7-6 所示。

表 7-6　测试集手工标注句子结果

测　试　集	主题句数目	褒义主题句数目	贬义主题句数目
T_1	1 023	351	278
T_2	1 437	698	362
T	2 702	1 129	597

使用本方法在测试集 T_1、T_2 和 T 上分别进行测试，表 7-7 为各次实验结果。

表 7-7　主题句提取的实验结果

实验测试集	P	R	F_1
T_1	0.731	0.744	0.742
T_2	0.744	0.765	0.754
T	0.743	0.778	0.760
平均	0.739	0.766	0.752

实验结果表明，本方法的主题句识别平均准确率 P 达到 73.9%，平均召回率 R 达到 76.6%，平均 F_1 值也达到了 75.2%，算法的整体性能比较高。

2．算法性能对比

为了验证本方法的主题句识别性能，现将本方法与 SVM 方法进行对比实验，比较两种方法的识别性能。

首先采用本方法识别出文本的主题句并标注句子的褒贬倾向，并将其结果直接应用于 SVM 方法中。然后分别使用本方法和 SVM 方法对文本的情感进行分析，图 7-2 为两种方法的实验结果。

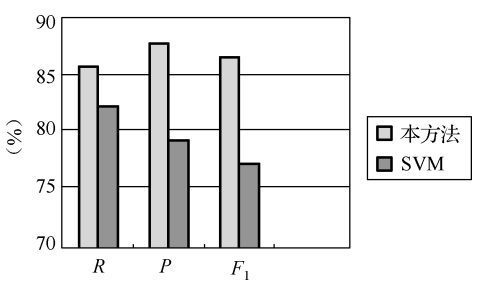


图 7-2　两种算法性能对比

从图 7-2 可以看出，本方法的各项性能指标都优于 SVM 方法，特别是准确率 P 和 F_1 值，分别比 SVM 方法高出约 10%和 11%，达到了较高的性能水平。

3．句子主观关系识别算法

1）实验数据集

实验数据集是一个关于手机产品评论的中文语料，共有 1 600 篇文本，数据集分为两部分，1 200 篇作为训练语料集，400 篇作为测试语料集。

为了准确地评价算法的性能，从测试语料库 T 中随机选取 30%、60%的测试语料，形成两个子测试集 T_1 和 T_2，实验分别在 T_1、T_2 和 T 上进行。

2）算法性能验证

在实验中，使用最大熵算法来提取句子中的主观关系。将基本最大熵算法（简称 Baseline 方法）与结合了基本特征和语义特征的 Baseline 方法进行比较，考察基本特征和语义特征对句子主观关系识别性能的影响，评价指标采用准确率（P）、召回率（R）和 F_1 值。

（1）Baseline 方法实验

首先在 T_1、T_2 和 T 三个语料集上对 Baseline 方法进行实验，其实验结果如表 7-8 所示。

表 7-8　Baseline 方法的实验结果

语　料　集	P	R	F_1
T_1	0.386	0.778	0.516
T_2	0.390	0.777	0.519
T	0.396	0.777	0.524

从表 7-8 中的实验数据可以看出，召回率 R 指标比较好，而准确率 P 和 F_1 值指标比较差。在实际应用中，相对于召回率，准确率更为重要。因此，Baseline 方法的识别性能比较低。

（2）各项基本特征对性能影响实验

在下面的实验中，所有的实验结果都是在 T_1、T_2 和 T 三个测试语料集上经过多次实验得到的平均值。为了考察各项基本特征对识别性能的影响，在 Baseline 方法上分别增加词、词性和距离等特征，其实验结果如表 7-9 所示。

表 7-9　各项基本特征对识别性能的影响

特　　征	P	R	F_1
Baseline	0.390	0.777	0.519
Baseline+词	0.647	0.635	0.641
Baseline+词性	0.602	0.623	0.612
Baseline+距离	0.553	0.678	0.609

从表 7-9 中的实验数据可以看出，Baseline 方法分别增加了词、词性和距离等特征后，

其准确率 P 和 F_1 值都有不同程度的提高，其中词和词性特征对性能提高的作用比较明显，这是因为词本身蕴涵的信息非常丰富，很多词本身就带有明确的倾向性，如“好”、“漂亮”等。词性也是如此，情感词是最常见的形容词性，很多词的词性也具有明确的指向性。距离特征对提高性能的作用也比较大，虽然略低于前两个特征。实验结果表明，各项基本特征对提高识别性能都起到了较大的作用。

（3）基本特征综合对性能影响的实验

下面的实验是考察分别将词与词性特征两种特征以及词、词性和距离三种特征综合起来对识别性能的影响，其实验结果如表 7-10 所示。

从表 7-10 中的实验数据可以看出，在 Baseline+词的基础上，再增加词性和距离特征后，其准确率 P、召回率 R 和 F_1 值都有不同程度的提高。实验结果表明，将各项基本特征综合起来应用，能够进一步提高识别性能，但提高的幅度有限。

表 7-10　基本特征综合对性能的影响

特　征	P	R	F_1
Baseline+词	0.647	0.635	0.641
Baseline+词+词性	0.649	0.636	0.642
Baseline+词+词性+距离	0.653	0.637	0.645

（4）语义特征对性能影响的实验

为了考察语义特征对识别性能的影响，在 Baseline+基本特征上增加语义特征，其实验结果如表 7-11 所示。

表 7-11　语义特征对性能的影响

方　法	P	F_1
Baseline 方法	0.390	0.519
Baseline+基本特征	0.653	0.645
Baseline+基本特征+语义特征	0.712	0.683

从表 7-11 中的实验数据可以看出，在 Baseline+基本特征的基础上，再增加语义特征后，其准确率 P 和 F_1 值都有较大的提高。实验数据表明，在提取句子中的主观关系时，充分考虑评价词和评价对象本身的词信息，尤其在模型训练中加入语义信息，能够更准确地识别出主观句中的主观关系。

7.4　段落情感分析方法

段落情感分析的对象是经过文本分割后的语义段而不是自然段落。由于语义段之间存在着语义联系，因此有助于对文本情感进行细化分析。语义段分割方法的详细介绍见 6.4 节。

在语义段情感分析时，以语义段中的句子为基本单元，通过计算句子情感值和语义段情感值，最终得到文本的全局情感值，实现对整个文本的情感分析。

下面介绍语义段句子情感、语义段情感以及文本全局情感的分析方法。

7.4.1　语义段句子情感标注

一个语义段由多个句子组成，各个句子的情感对该语义段的情感起到重要的作用，因此要确定一个语义段的情感，必须首先确定该语义段中各个句子的情感。

除了句子的情感对语义段的情感产生影响外，句子的重要程度也会影响到语义段的情感。因为每个句子在文本中的作用是不同的，重要句子不仅反映了文本的内容，并且还起到与其他文本相区分的作用，而次要句子的这种作用则比较弱。因此，使用句子权重来表示一个语义段中各个句子的重要程度，根据每个句子的重要程度赋予不同的句子权重。

考虑到一个语义段的情感是由句子情感值和句子权重两方面因素来确定，可以用下式来计算一个语义段的情感倾向值：

$$P(C_i)=\sum_{0<j<n} w_{s_j}\, p(s_j) \qquad (7\text{-}11)$$

式中，$P(C_i)$表示语义段 C_i 的情感倾向值，n 表示该语义段包含的句子数量，w_{s_j} 表示句子 s_j 的权重，$p(s_j)$表示句子的情感倾向值。

为了确定句子的情感倾向，需要对语义段中的句子情感倾向（褒义或贬义）进行标注。由于一个语义段中包含了主观句子和客观句子，只有主观句子具有情感倾向，因此需要首先识别出语义段中的主观句子，然后对主观句子的褒贬倾向进行标注。主观句子识别和褒贬倾向标注方法见 7.3 节。

7.4.2　语义段句子权重计算

句子权重计算方法主要有三种，第一种方法认为每个段落的第一句或者最后一句具有较高的权重，这种方法主要适合于结构化或半结构化的文本。第二种方法认为标题是文本主要内容的提炼，因此标题具有较高的权重，缺点是如果标题不能充分反映文本的内容，则会增加分类的模糊性。第三种方法是结合了两种方法来计算句子权重：一是基于标题的方法，二是基于句子中各个特征项的方法。

1．基于标题的方法

基于标题的方法是通过计算句子与标题的相似度来确定该句子的权重，如果句子与标题具有很高的相似度，就赋予句子相对较高的权重，反之则赋予较低的权重。

将标题与句子表示成 VSM 模型中的特征项向量，向量间的夹角越小，标题与句子的相似度越高。句子与标题相似度的计算公式如下：

$$\mathrm{Sim}(S_i,T)=\frac{S_i\times T}{\|S_i\|\times\|T\|} \tag{7-12}$$

式中，T 代表标题的向量，S_i 代表句子的向量。需要设置相似度阈值 δ，该值由实验来确定，根据多次实验结果，设置 $\delta = 0.75$ 比较合适。当一个句子与标题的相似度大于阈值时，则认为该句子与标题相似，同时认为包含该句子的语义段相对比较重要。用式（7-12）可以计算获得基于标题的统计结果。

2．基于特征项的方法

首先需要衡量句子中每个特征项的重要性，然后根据特征项计算句子的权重。

由于词是构成句子、语义段以及文本的基本元素，从文本的所有词中抽取出能够反映文本特征的词构成文本的特征项，并按某一方法赋予特征项相应的权重。特征项的权重综合反映了该特征项对文本内容的贡献度和对不同文本内容的区分能力。特征项在不同文本中出现的频率满足一定的统计规律，因此可以通过特征项的频率特性来计算其权重。一个有效的特征项集合应具有以下两个特征：

（1）完全性：特征项能够反映目标文本的内容；

（2）区分性：特征项具有将目标文本和其他文本相区分的能力。

根据以上两个特征，特征项权重计算应满足两个原则：一是正比于特征项在文本中出现的频率；二是反比于文本集合中出现该特征项的文本频率。特征项权重计算方法主要有布尔函数、TF-IDF 函数等方法，常用的是 TF-IDF 函数方法。

在 TF-IDF 函数方法中，首先需要统计特征项的 TF 和 IDF 值，然后把句子中每个特征项的 TF 和 IDF 的乘积值进行加权并归一化处理，获得基于特征项的统计结果，其计算公式如下：

$$I(s_j)=\frac{\sum_{t\in s}\mathrm{tf}(t)\times\mathrm{idf}(t)}{\max_{s\in d}\left[\sum_{t\in s}\mathrm{tf}(t)\times\mathrm{idf}(t)\right]} \tag{7-13}$$

式中，$tf(t)$表示词的频率，idf(t)表示倒转文本频率，d 表示文本，s 表示句子。

结合上述两种统计方法的计算结果，句子的总权重计算公式如下：

$$w(s_j)=1+I(s_j)+\mathrm{Sim}(S_j,T) \tag{7-14}$$

7.4.3 语义段情感计算方法

在完成语义段的句子情感标注和句子权重计算后，根据式（7-11），将语义段中的每个句子的情感值与权值相乘并求和，最终得到该语义段的情感倾向值。

一个语义段只是文本的一部分，语义段的情感仅代表了一个文本的局部情感，在此基

础上，还需要确定整个文本的情感，也称为文本的全局情感。全局情感可以看作一个基于局部情感的函数，局部情感作为该函数的变量，对全局情感的最终倾向性产生重要影响，可以认为文本中所有局部情感的互相影响产生了文本的全局情感。对于全局情感，可以采用加权求和方法、KNN 算法等来计算。

1．加权求和方法

加权求和方法是最简单的方法，通过对各个语义段的情感值加权求和来确定全局情感值。

根据语义段在文本中的位置不同，各个语义段的情感对全局情感的影响因子也不同，按照文本的通常写作习惯，文本的首尾部分比其他部分的内容更为重要，具有提示和概括的作用。因此，在确定文本的全局情感时，首尾部分的局部情感权重要大于其他部分。

在计算出每一个语义段的情感倾向值后，按照下式计算文本的全局情感值：

$$T = \sum_{i=1}^{k} C_i \times Q \tag{7-15}$$

式中，T 为全局情感值，Q 是语义段 C_i 的情感权值，根据语义段在文本中位置的不同，赋予其不同的权值，最简单的方法是对包含首段或尾段的语义段赋予大于 1 的权值，而其他语义段的权值均赋为 1，这种方法也称为 Q 值方法。

2．KNN 算法

在 5.5.1 节中介绍了 KNN 算法及其在文本分类中的应用。下面应用 KNN 算法来计算文本的全局情感值。

基于 KNN 算法的全局情感值计算公式如下：

$$P(T, s_j) = \sum_{i=1}^{k} P_T(C_i, s_j) \times W_i \tag{7-16}$$

式中，$P(T, s_j)$表示文本 T 的全局情感值；$P_T(C_i\ s_j)$表示文本 T 中语义段 C_i 的情感值，即文本的局部情感值，当语义段 C_i 中的句子 s_j 为褒义时，$P_T(C_i\ s_j)$取值为 1，当语义段 C_i 的句子 s_j 为贬义时，$P_T(C_i\ s_j)$取值为–1；W_i 表示语义段 C_i 的权值。

语义段权值的计算是确定全局情感的关键步骤。由于一个文本中的每个语义段对文本的贡献大小是不同的，因此采用语义段贡献率来定义语义段的权值，贡献率是指语义段对文本主题及内容的贡献程度。根据对语料库中文本的统计，对一个语义段的贡献率产生影响的因素主要有以下几个方面：

（1）语义段中包含主题句的数量。如果一个语义段包含的主题句相对较多，则说明该语义段对文本主题表达发挥的作用更大，因此包含越多主题句的语义段的贡献率相对越大。各个语义段中包含的句子的数量也会对这一因素产生影响，当包含主题句数量相同时，语义

段本身的句子数量越小，其贡献率则相对较大。

（2）语义段是否包含文本的首段或者尾段。按照文本的语义结构和作者的写作习惯，文本的首段通常具有说明文本主题的作用，而尾段通常对文本表达的主题进行总结或强调，可见首段和尾段对文本表达更为重要。因此，包含文本首段或者尾段的语义段具有较大的贡献率。

（3）语义段是否包含与文本标题相似度较高的句子。文本的标题是对文本主要内容的提炼，具有较高的重要度。如果语义段包含了与标题相似度高的句子，则说明该语义段对文本内容的表达更为重要，因此也具有较大的贡献率。

综合考虑上述各个因素，一个语义段的贡献率定义如下：

$$V_{C_i}=1+\frac{S(C_i)}{N_{C_i}}+\lambda+\frac{\sum_{0<j\leqslant N_{Ci}}\mathrm{Sim}(s_j,T)}{n} \tag{7-17}$$

式中，V_{C_i} 表示第 i 个语义段的贡献率，N_{C_i} 表示语义段 C_i 中包含的所有句子的数量，$S(C_i)$表示语义段 C_i 中包含的主题句子的数量，λ 表示当语义段包含首段或者尾段时赋予的权值，n 表示语义段 C_i 中所包含的与标题相似的所有句子的数量。λ 值需要通过实验来确定。

通过式（7-17），得到一个语义段的贡献率，即语义段的权值。然后通过式（7-16）计算出一个文本的全局情感值。

7.4.4 算法验证

下面通过实验数据对语义段情感分析方法及其效果进行测试和验证。

1．语义段情感分析效果

1）实验数据集

实验数据集是一个关于手机产品评论的中文语料，选取针对苹果手机的评论文章 500 篇，其中 300 篇为训练文本，200 篇为测试文本，同时手工分割和标注每个文本的语义段。

2）语义段效果验证

下面的实验是考察基于语义段的情感分析效果，首先对测试文本集中的所有文本进行语义段分割，然后对每个文本中的各个语义段的主题和情感倾向进行判别。例如，一篇测试文本中包含有 6 个自然段落，经过分割后形成 3 个不同主题的语义段，对每个语义段的情感倾向进行判别，其实验结果如表 7-12 所示。

表 7-12　对一个测试文本的实验结果

语义段包含的自然段落	语义段的主题	语义段的情感倾向
第 1 和第 6 段	苹果手机	褒义
第 2 和第 3 段	外形	褒义
第 4 和第 5 段	电池	贬义

该测试文本包含了三个语义段，各个语义段的主题分别是“苹果手机”、“外形”和“电池”。主题为“苹果手机”的语义段包含了首尾自然段，是对手机的整体评价，其倾向性是褒义；而主题为“外形”的语义段包含了文本的第 2、3 自然段，其倾向性是褒义；主题为“电池”的语义段包含文本的第 4、5 自然段，其倾向性是贬义。从表 7-12 的实验结果可以看出，连续的自然段在语义上相近，通常属于一个语义段，例如第 2、3 自然段属于同一个语义段，第 4、5 自然段也是属于同一个语义段。

对每个测试文本进行上述的分析后，再对同一主题的情感倾向性进行赋值并求和，褒义赋值为 1，贬义赋值为–1，其实验结果如表 7-13 所示。

表 7-13　对测试集的实验结果

语义段主题	主题情感倾向	情　感　值
外形	褒义	1
音质	褒义	1
电池	贬义	-1
通话效果	褒义	1
拍照效果	褒义	1
安全性	褒义	1
价格	贬义	–1

表 7-13 中的主题是测试集中出现的所有主题，测试集中的每个文本不一定包含所有主题，可能只包含其中的若干个主题。表 7-13 包含了对苹果手机的各个方面评价。可见，基于语义段层次的情感分析能够细化对文本情感的分析，提高了文本情感分析的效果。

2．语义段情感分析方法

1）实验数据集

实验数据集仍然是一个关于手机产品评论的中文语料，共有 1 200 篇文本，对于语料中的每个文本，手工识别和标注每个语义段及语义段的情感倾向。数据集分为两部分，600 篇作为训练语料集，600 篇作为测试语料集。

2）算法有效性验证

式（7-11）给出了一种语义段情感计算方法，在计算语义段情感时考虑了句子情感和句子权重两方面因素，能够提高语义段情感分析的性能。为了验证该方法的有效性，对两种计算方法进行对比实验，其中方法 1 是没有考虑句子情感和句子权重的计算方法，方法 2 是同时考虑了句子情感和句子权重的计算方法，评价指标采用召回率（R）、准确率（P）和 F_1 值，其实验结果如表 7-14 所示。

表 7-14　两种语义段情感计算方法的实验结果

实验方法	R	P	F_1
方法 1	0.823	0.809	0.816
方法 2	0.857	0.896	0.876

从表 7-14 中的实验数据可以看出，方法 2 的各项性能指标都超过了方法 1，表明在同时考虑句子情感和句子权重两方面因素后，确实能够有效地提高语义段情感分析的性能。

3．算法性能验证

在计算语义段权值时，主要有两种方法：一种是 Q 值方法，该方法仅考虑到包含首段或者尾段的语义段有较大权值，而不考虑其他情况；另一种方法是贡献率方法，该方法考虑了标题、主题句子等多种因素对语义段权值的影响。通过两种方法的对比实验，考察哪种方法更加有效。

在实验前，首先需要确定 Q 值方法中的首尾段加权值 Q。设置 Q = 1.1, 1.2, …, 1.9，分别在 600 篇语料样本上进行试验，其中褒义倾向文本 350 篇，贬义倾向文本 250 篇，使用公式（7-11）来计算语义段情感值。随着 Q 选取不同的值，语义段情感识别准确率也随之发生变化，Q 值与准确率之间的关系如图 7-3 所示。从图 7-3 可以看出，当 Q = 1.6 时，情感识别准确率最高。因此在下面的实验中，Q 取值设为 1.6。

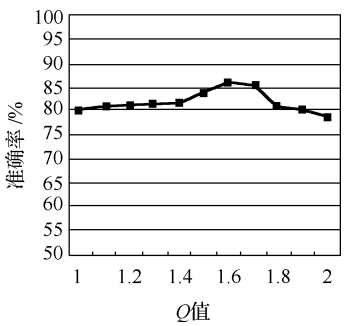


图 7-3　Q 值与准确率关系图

在 Q 值方法和贡献率方法的对比实验中，选取测试集文本数的 30%、50%、70%、80%、95%，构成 5 个测试集，分别记为 T_1、T_2、T_3、T_4 和 T_5。在 5 个测试集上，分别采用 Q 值方法和贡献率方法计算语义段权值，并且获得文本的全局情感，评价指标仍然采用准确率（P）、召回率（R）和 F_1 值，其实验结果如表 7-15 所示。

表 7-15　两种权值计算方法的实验结果

测试集	Q 值方法			贡献率方法		
	P	R	F_1	P	R	F_1
T_1	0.805	0.821	0.813	0.849	0.831	0.840
T_2	0.813	0.819	0.816	0.841	0.835	0.837
T_3	0.824	0.820	0.823	0.853	0.840	0.846
T_4	0.832	0.825	0.828	0.855	0.842	0.8.8
T_5	0.847	0.823	0.835	0.860	0.845	0.852

从表 7-15 中的实验数据可以看出，贡献率方法的各项性能指标都超过了 Q 值方法，表明采用贡献率方法来计算语义段权值，能够有效地提高文本情感识别的性能。

4．算法性能对比

在文本的全局情感分析中使用了 KNN 算法。为了验证 KNN 算法的全局情感识别性能，现将 SVM 和贝叶斯（Bayes）两种常用的分类算法与 KNN 算法进行对比实验，在 5 个测试集 T_1、T_2、T_3、T_4 和 T_5 上，分别采用三种算法来计算和识别文本的全局情感，评价指标仍然采用召回率（R）、准确率（P）和 F_1 值，其实验结果如表 7-16 所示。

表 7-16　三种算法的实验结果

测试集	SVM			Bayes			KNN		
	P	R	F_1	P	R	F_1	P	R	F_1
T_1	0.808	0.813	0.810	0.805	0.791	0.798	0.885	0.861	0.873
T_2	0.813	0.816	0.814	0.806	0.798	0.802	0.889	0.873	0.881
T_3	0.819	0.820	0.819	0.810	0.802	0.806	0.892	0.876	0.884
T_4	0.821	0.823	0.823	0.812	0.808	0.810	0.895	0.879	0.887
T_5	0.827	0.825	0.826	0.817	0.816	0.816	0.902	0.885	0.893

从表 7-16 中的实验数据可以看出，在三种算法中，KNN 算法的各项性能指标都是最好的，达到了比较高的性能水平。因此，在文本的全局情感分析中使用 KNN 算法是比较适宜的。

7.5　文本情感分析模型

网络舆情分析的对象是网络文本，网络文本中的语言表达方式有其自己的特点，采用语言建模方法对网络文本的情感倾向进行分析，有助于提高对网络文本情感分析的效果。在语言建模中，首先选择一种统计类语言模型作为基本语言模型，然后在标注有褒贬倾向的训练文本集上对情感模型进行估计。对于每一个测试文本，比较其语言模型与情感模型之间的相似度，如果与某个情感模型更为相似，则认为该文本的褒贬倾向与这个模型的褒贬倾向相一致，从而实现对文本情感倾向的识别。

下面介绍基于语言建模的文本情感分析方法。

7.5.1　文本情感模型

人们在使用语言表达不同的情感时，一方面要遵循一定的语言表达规则，让他人能够明白其表达的内容；另一方面每个人都有不同的语言表达习惯，即因人而异的语言特色。因此，根据语言表达规则和习惯差异性，可以通过语言建模方法对文本情感进行分析。

1．统计语言模型

统计语言模型是基于统计方法的自然语言处理模型，使用分布函数来表示词、词组及句子等自然语言的基本单位，通过统计方法来描述自然语言的生成和处理规则。统计语言建模（SLM）是常用的语言建模技术，SLM 认为，任何语言表达实质上就是其字母表上的某种概率分布，该分布反映了任何一个字母序列成为该语言表达的一个句子的可能性，这个概率分布称为语言模型。

对于任何一个句子 S，其概率分布计算公式如下：

$$P(S)=\prod_{i=1}^{l}P(w_i\,|w_1w_2\cdots w_{i-1}) \tag{7-18}$$

式中，w_i 表示词语。SLM 的关键问题是如何根据给定的语料库来估计概率 $P(w_i|w_1w_2\cdots w_{i-1})$。由于在实际中不可能有足够多的数据来估计 $P(w_i|w_1w_2\cdots w_{i-1})$，因此通常采用 N 元模型来估计，N 元模型认为一个词的出现与否仅与其前面的 N–1 个词相关，即 $P(w_i|w_1w_2\cdots w_{i-1})\approx P(w_i|w_{i-n+1}\cdots w_{i-1})$。当 $N=3$ 时，称为三元模型（trigram）；当 $N=2$ 时，称为二元模型（bigram）；当 $N=1$ 时，称为一元模型（unigram），此时假设各个词之间是相互独立的。

在实际应用中，需要根据不同的应用需求选取不同的 N 元模型。例如，在语音识别或机器翻译等应用中，词序的作用比较大，通常使用 trigram 模型等高阶模型；在信息检索、文本分类等应用中，词序的作用相对较小，一般采用 unigram 模型或 bigram 模型。在情感分析建模中，使用 unigram 和 bigram 两种模型。

为了建立 N 元模型，概率估计主要基于训练语料中 N-gram 模型的出现频率。例如，在二元模型中，对 bigram 模型的出现频率进行估计；在一元模型中，对 unigram 模型的出现频率进行估计。在对模型参数进行估计时，即使语料库规模很大，也会有很多可能的 N-gram 模型没有出现在语料库中，这种情况将会影响统计结果的准确性。因此，在建立 N 元模型后，必须做数据平滑处理。

在建立 N 元模型时，模型阶数 N 和建模单元是影响模型性能的两个重要因素。在建模单元相同的情况下，高阶模型的性能要优于低阶模型，但高阶模型的构造难度要大于低阶模型。理论上，N 值越大，所反映的语序也越逼近于真实的句法模式，语法匹配效果更好，模型更加逼近真实的语言现象。但是，在实际应用中，N 值的增大又会带来存储资源的急剧增加以及因统计数据稀疏而造成的计算误差。

建模单元是影响模型性能的另一个重要因素，在模型阶数 N 相同的情况下，基于词的模型要优于基于字的模型，但建模的复杂度相对大一些。

为了比较待处理文本与模型之间的相似度，通常采用 KL（Kullback-Liebler）距离来度量，KL 距离也称为交叉熵，常用于度量两个正函数之间的相似度，对于两个完全相同的函数，它们的交叉熵等于 0。在自然语言处理中，使用 KL 距离来度量两个词在语法和语义上

是否同义，或者两篇文章的内容是否相近。

两个概率分布 $q(x)$与 $p(x)$之间的 KL 距离定义如下：

$$\mathrm{KL}(q(x)\|p(x))=\int q(x)\ln\left[\frac{q(x)}{p(x)}\right]\mathrm{d}x \tag{7-19}$$

$q(x)$与 $p(x)$之间差别越大，KL 距离越大；当 $q(x)$与 $p(x)$完全相同时，KL 距离最小，其值为 0。

基于上述的统计语言模型，对文本中所表达的情感进行模型化。语言表达的模型化实际上是对某种语言单元在文本中分布概率的统计，表明观察到该类语言单元出现的可能性。因此，情感模型就是对情感语言单元在文本中分布概率的统计。

在建立情感模型之前，首先需要构建一个基本的语言模型，其过程步骤如下：

（1）基本语言模型需要针对文本语料来构建，模型的构建过程如下：设 X 是文本的集合，$X=\{x_1, x_2, \cdots, x_n\}$，$C$ 表示文本类别的集合，是对 X 的一个划分：$C=\{c_1, c_2, \cdots, c_k\}$，$c_i \cup c_j=\varphi$，$\forall i \neq j$。

（2）使用 KL 距离对模型进行估计，通过密度函数来定义 KL 距离。在文本集合 X 中，x 的概率分布定义如下：

$$p(x)=\sum_{i=1}^{K} p(x|c_i)\times p(c_i) \tag{7-20}$$

式中，数据 x 在第 i 类上的概率密度函数为 $q(x) = p(x|c_i)$。密度函数 $p(x)$与 $q(x)$之间的 KL 距离定义如下：

$$\Psi=-KL(p(x|c_i)\|p(x)) \tag{7-21}$$

根据构建的模型不同，i 的取值不同。

2．情感模型构建

情感模型是对文本中所表达的情感倾向进行语言建模。在构建情感模型时，首先需要手工标注训练文本集中的文本情感，分别对褒义模型和贬义模型两种情感模型进行估计。然后分别计算待处理文本自身的语言模型与情感模型之间的相似度，如果待处理文本自身的语言模型与某个情感模型更为相似，则认为该文本的褒贬倾向与这个模型的褒贬倾向相一致。

在基本语言模型的基础上，式（7-21）中的 i 值为 2，表示有两种情感模型：褒义模型和贬义模型。情感模型的距离函数定义如下：

$$\theta(t,\delta_P,\delta_N)=d_1-d_2 \tag{7-22}$$

式中，t 表示待处理文本，δ_P 和 δ_N 分别表示褒义模型和贬义模型，d_1 代表文本 t 与褒义模型之间的 KL 距离，d_2 代表文本 t 与贬义模型之间的 KL 距离。当 θ 大于 0 时，表示待处理文

本更接近于贬义模型，判断文本表达的情感为贬斥类；当 θ 小于 0 时，判断文本表达的情感为褒扬类。当 θ 等于 0 时，表示文本表达的情感为中立。

7.5.2 模型参数估计

语言模型建立后，需要从语料中获取模型参数。模型参数是指使用该模型分析语言时所需要的统计数据。模型参数一般要利用训练语料来确定。确定语言模型参数的过程可以看作一个机器学习的过程。如果语言模型中的各个随机变量是词，对于 unigram 模型需要统计出每个词的出现频率，对于 bigram 则需要统计出每两个词的转移概率，对于 trigram 则需要统计出每三个词的转移概率（从任意两个词到任意的第三个词的转移概率）。

获取模型参数主要有两种方法：有指导的参数求解和无指导的参数求解。有指导的参数求解是指在人工标注的熟语料库的情况下进行机器学习的过程；而无指导参数求解是指在没有熟语料情况下的自学习过程。下面采用的是有指导的参数求解方法。

1．最大似然参数估计

由于语言模型中采用词的 unigram 和 bigram 作为模型参数，因此首先需要对参数的分布进行概率估计。在一个具有一定规模的训练语料集上，通常采用基于词频统计的最大似然估计（Maximum Likelihood Estimate，MLE）方法对两种参数进行估计。该方法使用样本的相对频率作为其真实概率分布的近似，设一个事件 E 出现 r 次，所有可能事件的出现次数总和为 N，最大似然估计的公式为：$p(E) = r/N$。

应用 MLE 方法对模型参数进行初步估计，其计算公式如下：

$$P_M(w_i|T) = \frac{\text{count}(w_i)}{\text{count}(r)} \tag{7-23}$$

式中，T 既可以表示待处理文本，也可以代表褒义文本集合或贬义文本集合。count(w_i)表示 unigram 或者 bigram 在 T 中出现的次数，count(r)表示任意一个词在 T 中出现的次数。

2．数据平滑技术

在语言模型估计中普遍存在数据稀疏问题。对于一个确定的训练语料，即使规模相当大，也会有大量的词串没有同时出现，这样就会出现大量估计值为零的情况，这就是所谓的数据稀疏问题。

在数据稀疏的情况下，使用 MLE 方法来估计不出现或者出现次数很少的事件的概率是不可靠的，必然会引起零概率问题：对于某个没有出现在文本 t 中的词项 w，使用 MLE 方法将导致 $P(w \mid t) = 0$。零概率问题会大大削弱模型的描述能力和处理能力。另一方面，由于语料库中噪声信息的干扰，按照 MLE 方法估计出的概率对真实概率分布的

近似可能存在较大的偏差。因此需要对 MLE 方法的概率进行调整，这种概率调整技术称为数据平滑技术。

数据平滑技术通过适当调整概率分布的取值，使低概率（包括零概率）被调高，高概率被调低，从而避免了零概率的出现，能有效地解决数据稀疏问题，同时还能使模型参数概率分布更加均匀，概率的计算更加精确。常用的数据平滑方法概括如下：

$$P(w_i|T)=\begin{cases}P_s(w_i|T), & \text{if } w_i \text{ is seen}\\ \alpha_d P_{ml}(w_i|C), & \text{othersize}\end{cases} \tag{7-24}$$

式中，$P_s(w_i|T)$表示在文本或文本集合 T 中可以观察到的语言单位 w_i 平滑概率，$P_{ml}(w_i|C)$是 w_i 在整个语料库 C 中出现概率的最大似然估计，α_d 是对没有出现的语言单元分配概率值时的控制权值。

为了获得更好的参数估计效果，对 unigram 模型和 bigram 模型分别采用不同的参数估计方法。

（1）对于 unigram 模型，比较适合采用 Dirichlet prior 平滑方法。Dirichlet prior 平滑是一种线性插值的方法，主要用于解决训练文本集较小时参数估计中的偏置问题：从可以观察的 N-gram 中分配适当的折扣（非零折扣）给文本集中没有观察到的 N-gram。针对 unigram 模型参数的平滑估计公式定义如下：

$$P_s(w_i|T)=\frac{\text{count}(w_i)+\alpha P_M(w_i|C)}{\text{count}(r)+\alpha} \tag{7-25}$$

（2）对于 bigram 模型，比较适合采用 Jelinek-Mercer 平滑方法，Jelinek-Mercer 平滑方法也是一种线性插值的方法，采用绝对折扣和插值模型相结合的方法，常用于解决由于训练样本集较小而引起的参数估计偏置问题。针对 bigram 模型参数的平滑估计公式定义如下：

$$P_s(w_i|T)=\lambda P_M(w_i|T)+(1-\lambda)P_M(w_i|C) \tag{7-26}$$

式中，λ 是一个平滑参数，$0<\lambda<1$，λ 直接影响模型的性能，需要通过实验来确定。

通过式（7-25）和式（7-26），完成对情感模型参数的估计和平滑。

7.5.3　语言模型评价

语言模型的性能反映了模型描述自然语言的能力，通过对语言模型的评价，可以了解语言模型性能的优劣，为语言建模提供指导。因此，语言模型评价是语言建模中一个很重要的问题。语言模型的性能可以采用两种评价方法：间接评价和直接评价，间接评价是指将语言模型嵌入语言处理系统中，通过语言处理系统的性能表现来评估其语言模型的优劣，由于语言处理系统涉及的处理任务比较多，各个任务之间互相影响，给语言模

型评价带来很大的复杂性。直接评价是指采用更为直接的评价方法和指标来评价语言模型的性能。

针对各种语言模型建立统一的评价指标是非常复杂和困难的，通常是针对特定的语言模型采用相应的评价指标。例如，针对 N-gram 模型，采用复杂度指标来评价。

语言模型的复杂度是衡量模型性能的一个重要指标。对于一个给定的语言模型，如果复杂度越小，则表明该语言模型越接近于客观存在的语言模型，并且更逼近于真实的语言现象，模型的质量也就越好。

常用的语言模型复杂度度量是根据模型计算出的测试数据概率或者利用交叉熵和困惑度等派生测度，模型复杂度的极小值为语言自身的熵。

1．交叉熵

利用在测试语料集上计算出的交叉熵作为模型性能的评价指标，交叉熵越小，说明模型与真实模型的差别越小，模型的性能表现越好。反之，交叉熵越大，说明模型与真实模型的差别越大，模型的性能表现越差。

对于一个随机变量 $X\sim p(x)$，$q(x)$是用于近似 $p(x)$的概率分布，那么随机变量 X 和模型 q 之间的交叉熵定义如下：

$$H(x,q)=H(x)+D(p\,\|\,q)=-\sum_{x}p(x)\lg q(x)=E_p\left(\lg\frac{1}{q(x)}\right) \tag{7-27}$$

因此，语言 $L=(X_i)\sim p(x)$与其模型 q 交叉熵定义为：

$$H(L,q)=-\lim_{n\to\infty}\frac{1}{n}\sum_{x_1^n}p(x_l^n)\times\lg q(x_l^n) \tag{7-28}$$

式中，$x_l^n=x_1,x_2,\cdots,x_n$ 为 L 的语句，$p(x_l^n)$ 为 L 中 x_l^n 的概率，$q(x_l^n)$ 为模型 q 对 x_l^n 的概率估计。虽然并不知道真实概率 $p(x_l^n)$，但是可以假设这种语言是“理想”的，并且假设语言 L 是稳态遍历的随机过程。因此，可以根据模型 q 和一个含有大量数据的 L 样本来计算交叉熵。在 n 足够大时，L 与其模型 q 的交叉熵计算公式可以简化为：

$$H(L,q)=-\frac{1}{n}\lg q(x_l^n) \tag{7-29}$$

通常，模型的交叉熵越小，模型的性能表现就越好。

2．困惑度

困惑度是另一个语言模型性能评价指标。给定语言 L 的样本 $l_l^n=l_1,\cdots,l_n$，则 L 的困惑度 PP_q 定义如下：

$$\mathrm{PP}_q = 2^{H(l,q)} \approx 2^{-\frac{1}{n}\lg(l_l^n)} = [q(l_l^n)]^{-\frac{1}{n}} \tag{7-30}$$

对于一个经过平滑操作的 N 元语法模型，其概率为 $p(w_i \mid w_{i-n+1}^{i-1})$。首先计算句子 $p(s)$概率，其计算公式如下：

$$p(s) = \sum_{i=1}^{l+1} p(w_i \mid w_{i-n+1}^{j-1}) \tag{7-31}$$

然后计算交叉熵，其计算公式如下：

$$H_p(T) = -\frac{1}{W_T}\lg_2 p(T) \tag{7-32}$$

$$p(T) = \sum_{i=1}^{l_T} p(t_i) \tag{7-33}$$

句子 $(t_1, t_2, \cdots, t_{l_T})$ 构成测试集 T，W_T 是以词为单位度量的文本 T 的长度，模型 p 的困惑度是模型分配给测试集 T 中每个词的概率几何平均值的倒数，它和交叉熵的关系为：

$$\mathrm{PP}_T(T) = 2^{H_p(T)} \tag{7-34}$$

通过式（7-32）、式（7-33）和式（7-34）的计算，可以得到语言模型的困惑度，困惑度越小，模型的性能越好。

7.5.4　算法验证

下面通过实验数据对文本情感分析模型及算法性能进行测试和验证。

1．实验数据集

实验数据是从互联网下载的 3 000 篇文章作为语料集，并手工标注所有文本的褒贬倾向，从语料集中选取 1 350 篇文章作为测试语料集。

2．模型复杂度评估

随机选用测试语料集中的 900 篇作为测试数据集，对 bigram 模型进行测试，考察该模型的最优模型复杂度。对于 bigram 模型，由于模型本身是以词为单位给出的模型概率，因此可以直接在词级插值。当模型中的插值系数取不同值时，模型复杂度变化情况如表 7-17 所示。实验结果表明，当插值系数 $\lambda = 0.4$ 时，模型的复杂度最低。

表 7-17　不同插值系数下 bigram 模型复杂度变化

λ	0.1	0.2	0.4	0.6	0.8	0.9
PPL	273.1	248.6	243.3	247.9	262.5	312

对于 unigram 模型，当 α 参数选取不同的值时，模型复杂度随之发生相应的变化，模型复杂度变化情况如表 7-18 所示。实验结果表明，当参数 α=450 时，模型的复杂度最低。

表 7-18　不同系数下 unigram 模型复杂度变化

α	100	250	350	450	550	750
PPL	234.1	220.7	215.9	199.8	213.5	228

在下面的实验中，λ 和 α 参数值分别取值为 0.4 和 450。

3．算法性能对比

采用准确率（P）和召回率（R）来评价算法的情感倾向识别性能，对比算法为 MLE 方法和 SVM 方法，其中，MLE 方法分为在参数估计中未做和做过数据平滑处理两种形式，SVM 方法采用 linear 作为核函数。表 7-19 是 MLE 方法与 SVM 方法的平均准确率，实验数据都是在不同特征集上通过 4 倍交叉验证得到的平均准确率。

表 7-19　MLE 方法与 SVM 方法的平均准确率

特　征	特征数目	平均准确率 (MLE)	平均准确率 (MLE+平滑)	平均准确率 (SVM)
bigram	10 912	0.813	0.885	0.779
unigram	5 697	0.631	0.738	0.729

表 7-19 的实验结果表明，直接使用 MLE 方法对模型进行参数估计，准确率不太理想，尤其对 unigram 模型的准确率比较低。在 MLE 方法的基础上经过数据平滑处理后，准确率有明显的提高。

对于 bigram 模型，MLE+平滑方法的准确率比 MLE 方法高出 7.2%，比 SVM 方法高出 10.6%，效果比较明显；对于 unigram 模型，MLE+平滑方法的准确率比 MLE 方法高出 10.7%，并略高于 SVM 方法，整体效果不如 bigram 模型那样明显，其主要原因是在模拟自然语言现象时，低阶模型（unigram）与高阶模型（bigram）之间存在一定的差距，因为越高阶的模型，对语言的模拟越逼真。另外，在语料规模不大的情况下，对 unigram 模型的数据平滑，有时会产生有噪声作用的特征。

4．算法性能鲁棒性对比

通常，训练集规模的变化会对算法的性能产生影响。为了全面评估算法的性能，通过变换训练语料集的规模，对 MLE 方法和 SVM 方法的性能鲁棒性进行测试和对比。

分别选取语料集的 10%、30%、50%、75%和 90%，形成 5 个不同规模的训练集，依次记为 T_1、T_2、T_3、T_4、T_5。测试集依然采用在前面的实验中所用的测试集。

在训练集规模不断增长，而测试集规模保持不变的情况下，MLE 方法与 SVM 方法的平均准确率如图 7-4 所示。

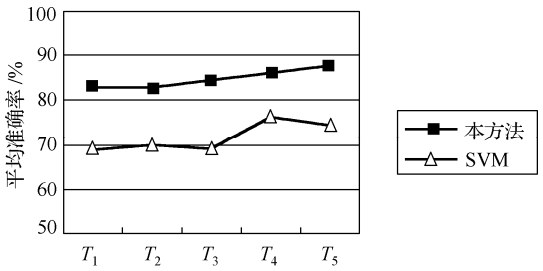


图 7-4　在不同规模训练集上的算法性能

图 7-4 的结果表明，对于 bigram 模型，MLE 方法随着训练集规模的增加，性能越来越好。SVM 方法也显示了这种趋势，但出现了不稳定情况。因此，与 SVM 方法相比，MLE 方法的鲁棒性更好。

参考文献

[1] 王来华. 舆情研究概论: 理论、方法和现实热点[M]. 天津: 天津社会科学院出版社, 2003

[2] 刘毅. 网络舆情研究概论[M]. 天津: 天津人民出版社, 2007

[3] 张兆辉, 郭子建. 舆情信息工作理论与实务[M]. 沈阳: 辽宁大学出版社, 2006

[4] 刘鹏飞, 周亚琼. 网络舆情热点面对面：突发公共事件舆情案例库 2015[M]. 北京：新华出版社，2016

[5] 小兵章嘎. 2015 年中国互联网舆情研究报告. 求是网(http://www.qstheory.cn/)，2016.1

[6] 政务微博. 百度百科(http://http://baike.baidu.com/subview/5725316/5774880.html)

[7] 张博. 面向主题的网络蜘蛛技术研究及系统实现[D]. 西北工业大学, 2009.3

[8] 罗知林. 微博网络用户转发行为及预测模型研究[D]. 西北工业大学, 2014.10

[9] 彭冬. 社交网络意见领袖发掘技术研究及应用[D]. 西北工业大学, 2012.3

[10] 徐会杰. 面向网络论坛的虚假舆情检测与抑制算法研究[D]. 西北工业大学, 2016.4

[11] 陈桂茸, 蔡皖东等. 一种网络论坛水军账号快速检测算法[J]. 湖南大学学报（自然科学版），Vol.42, No.4, 2015.4

[12] 石磊. 话题检测与追踪关键技术研究与算法实现[D]. 西北工业大学, 2010.3

[13] Jiawei Han, Micheline Kamber 著；范明, 孟小峰等译. 数据挖掘概念与技术[M]. 北京：机械工业出版社, 2001

[14] 杨惠. 文本分割模型与关键技术研究[D]. 西北工业大学, 2010.3

[15] 樊娜. Web 文本情感分析模型与方法研究[D]. 西北工业大学, 2010.1

[16] 蔡皖东. 网络空间信息传播建模分析[M]. 北京：电子工业出版社, 2017